Evaluation Considerations of Synthetic Natural Language Datasets for Question Answering Applications

Chris Van Buren*, Xiaotong Jiang, Jieyu Lin, Sachin Gopal Wani, Ajay Dholakia, David Ellison

Lenovo, Infrastructure Solutions Group, Morrisville, NC, USA
cvanburen@lenovo.com

Abstract. The rapid advancements in transformer-based language models have opened new possibilities for applications such as question answering and synthetic dataset generation. However, generating high-quality synthetic datasets can be computationally expensive and require significant resources. Fine-tuning a language model on a synthetic dataset can lead to improved performance on context-dependent questions, but it is crucial to evaluate the quality of these datasets beforehand. This study explores the benchmarking of synthetic natural language datasets generated with open-source language models for question answering applications. We fine-tune the same base model on each dataset independently and assess its performance on common benchmarks for question answering with context. Additionally, we examine metrics of the synthetic datasets themselves to identify potential indicators of downstream benchmark performance. Our results show that measuring the semantic similarity between questions is an indicator of domain diversity in a synthetic question-and-answer dataset. We also demonstrate that short answer lengths may indicate low quality chain-of-thought answers. Overall, Llama 3 8B Instruct performs most consistently well, of the models considered, in synthetic dataset generation. Our findings provide valuable insights for practitioners seeking to develop effective synthetic datasets for language-based applications.

Keywords: Generative AI, Synthetic Data, Language Models, Retrieval Augmented Fine-tuning, Open-Source Models, Natural Language Processing, Machine Learning

1 Introduction

The capabilities of transformer-based language models are rapidly advancing. Today's language models are being used for advanced question answering and synthetic data generation, among many other tasks. Generative language models are often used in applications alongside logic that retrieves recent or proprietary, unseen data and adds this data to the input to the language model, attempting to ground the model's response. This is an area of active research and has many practical applications, such as using models to answer context-dependent questions with high accuracy.

Another practical application of language models is synthetic dataset generation. Historically, many natural language datasets have been collections of existing human-created text [Schler et al., 2006; Zhu et al., 2015], whereas custom text creation for

datasets was very expensive and time-consuming. Now, however, language models with advanced abilities to generalize patterns and follow instructions can quickly generate near-human-level natural language.

Synthetic dataset generation can be used to improve context-dependent question answering applications when a custom dataset is created that demonstrates context-dependent question answering explicitly and a language model is fine-tuned to learn this pattern. This process, known as retrieval-augmented fine-tuning (RAFT), has shown to improve question answering capabilities of language models, especially in domain-specific contexts, in research from [Zhang et al., 2024]. While [Zhang et al., 2024] defines the process for generating a synthetic dataset for context-dependent question answering, it does not discuss the evaluation of these datasets to ensure they have the desired contents and quality prior to using them in large language model fine-tuning. They demonstrate the quality of the synthetic dataset implicitly through the downstream improvements in question answering benchmarks from their fine-tuned models. However, for this process to be generalizable and scalable, future practitioners will need to verify their synthetic datasets' quality, preferably prior to fine-tuning.

While RAFT is an exciting and effective process, fine-tuning a language model that has billions of parameters can be compute intensive and expensive [Lin et al., 2024]. The dataset used in training a model is crucial to its performance. It would therefore be beneficial for practitioners to have an idea of the quality of a synthetic dataset before committing to fine-tuning with it. There is, however, not a great deal of research in benchmarking the quality of synthetic natural language. We conduct an experiment by generating synthetic datasets with several different open-source language models. We fine-tune the same base model on each of these datasets independently and measure the fine-tuned models' performances on common benchmarks for question answering with context. Finally, we examine metrics of the synthetic datasets themselves to identify any potential indicators for downstream benchmark performance.

The study highlights that while model size and format do not directly correlate with downstream performance, understanding model size, architecture, and benchmarked capabilities remains an important consideration in synthetic dataset quality. Additionally, the percentage of correctly formatted samples, although it does not impact performance directly, plays a practical role in compute efficiency during generation. We observed that question similarity varies based on the dataset context (e.g., HotPotQA or PubMedQA). Domain-specific contexts tend to yield more similar questions than general-domain contexts. Additionally, shorter answer lengths in synthetic datasets may negatively impact downstream performance, emphasizing the importance of answer quality and reasoning capability in synthetic datasets intended for fine-tuning. In general, we find that the Llama 3 8B Instruct large language model performs the best, of the models considered, as a generator of synthetic question and answer datasets, and we recommend its use in similar cases. These insights are valuable for enhancing synthetic data quality and fine-tuning models effectively.

This paper is organized as follows: Section 2 summarizes related works. Section 3 describes the methodology followed in carrying out the work reported in this paper. Section 4 presents the results of the experiments. Section 5 is a discussion of various

factors that influenced the results and ends with a summary of future research directions. Finally, Section 6 provides the conclusions.

2 Related Work

2.1 Retrieval-Augmented Generation (RAG)

Since the invention of attention-based transformers [Vaswani et al., 2017], language models' abilities have grown to include general-purpose language understanding and generation [Minaee et al., 2024]. LLM's consists of billions of parameters which makes retraining for a specific domain expensive [Samsi et al., 2023]. RAG extends the knowledge of an LLM to a specific domain or knowledge base without retraining which makes it a cost-effective approach to improving LLM output staying relevant to the context. RAG systems have improved performance across various NLP tasks, especially open-domain question answering (QA) [Izacard et al., 2023; Lewis et al., 2020]. While RAG generally improves on the performance of isolated LLMs, it can face challenges from the failure of retrieval systems to return relevant context or failure of the LLM to properly integrate context into its response [Gao et al., 2023]. Fine-tuning an LLM with examples that demonstrate how to handle a lack of relevant context and how to synthesize context into responses can improve the reliability of RAG, as shown in [Zhang et al., 2024] through a process called retrieval-augmented fine-tuning, or RAFT.

2.2 Retrieval-Augmented Fine-Tuning (RAFT)

RAFT combines RAG and fine-tuning for adapting LLMs to specific domains [Zhang et al., 2024; Soudani et al., 2024]. It addresses the challenges of fine-tuning LLMs like over-specialization while also improving in-domain RAG performance. RAFT provides a more focused approach for in-domain tasks by demonstrating effective handling of distractor documents and relevant context extraction, improving reasoning abilities [Zhang et al., 2024]. Earlier work focused on the testing domain being different than that of the training time [Lin et al., 2023; Xu et al., 2023; Wang et al., 2023; Zhang, Y. et al., 2024]. RAFT's reliance on synthetic data generation for its training set highlights the amplifying demand for generated synthetic data. [Zhang et al., 2024] does not describe any analysis or cleaning process for their synthetic dataset used in RAFT. Since fine-tuning a model is generally compute- and time-intensive, it is important to qualify a synthetic dataset before using it in fine-tuning. Therefore, we look for indicators of quality in a variety of synthetic datasets by comparing measurable features of datasets with the downstream performance of the models fine-tuned on those datasets.

2.3 Synthetic Data and its Generation Using Large Language Models

The rising needs of large corpus datasets for training LLMs coupled with the scarcity of high-quality, diverse, and privacy- and license-compliant data has necessitated the generation of synthetic data. The use of LLMs for generation of synthetic data is an alternative approach to collection and curation of high-quality training data [Li et al.,

2023]. In addition to generation of synthetic data for Computer Vision (CV) tasks [Karras et al., 2018; Zhang et al., 2021; He et al., 2022; Besnier et al., 2020; Zhang et al., 2021], generative models have been used to create synthetic data for NLP tasks as well [Wang et al., 2021; Ye at al., 2022; Kumar et al., 2020; Chung et al., 2023; Gupta et al., 2023; Xu et al., 2022; Chen et al., 2022]. Research suggests a surge in generating synthetic data from LLMs, marking a notable shift in generative AI applications. However, the effectiveness of LLM-generated synthetic data is inconsistent across different tasks [Li et al., 2023] and it creates a need for accurately benchmarking the quality of synthetically generated datasets.

2.4 Benchmarking Quality of Generated Synthetic Datasets

Recent works have shown various techniques to generate better synthetic data [Veselovsky et al., 2023; Gupta et al., 2023] and its challenges [Guo et al., 2024]. Benchmarking its effectiveness involves evaluating the performance of models trained/fine-tuned on these synthetic data. There have been multiple studies involving various metrics on which synthetic datasets are benchmarked including F-1 score and accuracy among others [Layeghy et al., 2021; Singh et al., 2024; Jin et al., 2022]. LLM generated synthetic data are benchmarked on downstream tasks that measure accuracy [Gupta et al., 2023] our paper delves into understanding if insights regarding quality of datasets and compute required can allude to downstream benchmark results on fine-tuned models using QA datasets across various domains [Yang et al., 2018; Jin et al. 2019]. We examine intrinsic features of synthetic datasets, including semantic similarity between samples and sample length. We also consider characteristics of the LLMs used to generate synthetic datasets and the ability of those models to generate data consistently in the desired format.

3 Methods

3.1 Introduction to Methods

The RAFT [Zhang et al., 2024] process starts with a set of context documents: this is a collection of unstructured text. The process then uses an LLM to generate a dataset from the context documents, which we will explain thoroughly in Section 3.3. This synthetic dataset of {context, question, answer} triplets is then used to finetune an LLM, training it to respond with answers, using the provided question and associated context. We experiment by using various LLMs to generate the synthetic dataset. We attempt to find qualitative indicators of synthetic dataset quality by comparing dataset metadata with the performance of an LLM finetuned on that dataset on standard benchmarks.

3.2 Context Documents and Benchmarks

RAFT has demonstrated performance gains on several context-dependent question-answering benchmarks [Zhang et al., 2024]. Of those, we use PubMedQA [Jin et al., 2019] and HotPotQA [Yang et al., 2018]. Each of these datasets consists of datapoints that contain context documents, along with associated questions and answers. PubMedQA is a biomedical question answering dataset that tests a model's ability to answer questions with yes/no/maybe using abstracts from published medical papers as context. HotPotQA is a diverse, explainable, multi-hop dataset consisting of 113k question-answer pairs with context from Wikipedia article excerpts. To construct our synthetic datasets, we use the context document collections from PubMedQA and from HotPotQA. By using these datasets, we include both domain-specific and diverse benchmarks.

3.3 Synthetic Dataset Generation Process

For both PubMedQA and HotPotQA, we take a subset of the context documents from the datasets. The two context document collections are then independently used in prompts to an LLM to generate a dataset used for retrieval-augmented fine-tuning.

We synthetically generate questions and answers by prompting an LLM with a context document and instructions to generate a question and answer using the information in the document.

Table 1. For synthetic dataset generation, the document parameters are based on those determined to generalize to high downstream performance in [Zhang et al, 2024].

Parameter	Value
Number of distractor documents in each datapoint	4
Fraction of datapoints to contain relevant context documents	0.8
Number of datapoints per context document	2

To construct a datapoint for the fine-tuning dataset, we add four random "distractor" documents, documents that are not relevant to the question. Using distractor documents in the training set prepares the fine-tuned LLM for future unseen questions, when the context retrieval system may provide irrelevant documents that it should ignore. Additionally, the relevant document is removed from 20% of datapoints, preparing the model for instances where the context retrieval fails to retrieve any useful context. These parameters were determined to yield the best performance in [Zhang et al., 2024] and are summarized in Table 1. In conclusion, each datapoint in the synthetic dataset used for fine-tuning contains inputs: distractor context documents, a single relevant context document (sometimes omitted), and a question; and an output: the answer to the question. An example of a datapoint's inputs is shown in Figure 1.

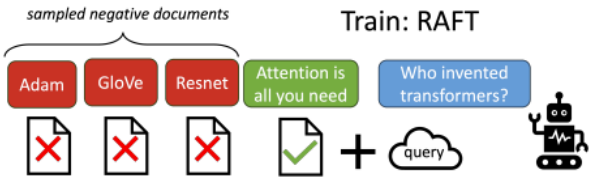

Figure 1. The components of a datapoint in a RAFT synthetic dataset are a question, a relevant context document, and other "distractor" documents, paired with an answer to the question. (Figure Source: Zhang et al., 2024.)

We use the following open-source LLMs to generate the datasets, using two model architectures and using model sizes spanning from 7 billion to 70 billion parameters:

- Llama 3 8B Instruct
- Llama 3 70B Instruct
- Llama 3 8B Instruct (q4_0 quantization)
- Mistral 7B (q4_0 quantization)

3.4 Synthetic Dataset Metrics

We evaluate several metrics about the generated synthetic datasets. These metrics consider the quality of the dataset and the amount of compute required to create that dataset.

Model Factors. We note the total count of parameters in each model used to generate datasets. We also note the format of the model parameters, such as the *bfloat 16* (*bf16*) floating point data type or *q4-0* 4-bit quantization.

Percentage of Correctly Formatted Synthetic Samples. We filter out samples from the synthetic dataset where the LLM did not follow the prompt or respond in the requested format. Specifically, for question generation, we verify if the generated string contains question words and has a length of fewer than 50 tokens. For chain-of-thought answer generation, we check whether the generated string contains the specified separator (i.e., "answer:")

Average Question Similarity. We measure the similarity of questions generated in each synthetic dataset by creating text embeddings with the all-MiniLM-L6-v2 embeddings model, then calculating the cosine similarity of each pair of question embeddings. We examine the distribution of cosine similarity scores, expecting a normal distribution. For each synthetic dataset, we report the average cosine similarity score, which serves as an indicator of the degree to which the generated questions are similar to one another, ranging from -1 (opposite) to 1 (identical), revealing the variety or uniformity in the topics and phrasing produced by the model.

Average Datapoint Length. We measure the average length (in characters) of each datapoint in the synthetic dataset.

Downstream Performance. We fine-tuned Llama 2 7B using each synthetic dataset and tested the models on the first 200 test samples from HotPotQA and PubMedQA, corresponding to the contexts used in the dataset generation. For each test sample, which includes a question and context as input, we used the fine-tuned model to generate an answer and measured its similarity to the ground truth answer.

For PubMedQA, we extracted "yes," "no," or "maybe" from the model output and calculated accuracy (exact match) and F1 score against the ground truth. For HotPotQA, we processed both the model output and the ground truth into bags of tokens and calculated accuracy (exact match) and F1 score at the token level for each sample. LLM Fine-tuning Process. We use the low-rank adaptation (LoRA) [Hu et al., 2021] technique to fine-tune the Llama 2 7B LLM. We use a learning rate of 0.0001, batch size of 128, and train for 50 steps.

4 Results

This section gives an overview of the results for each metric defined above. Table 2 and Table 3 show the synthetic dataset metrics and downstream benchmark performance for the HotPotQA and PubMedQA datasets, respectively.

Table 2. Synthetic dataset metrics and results on the HotPotQA benchmark.

Dataset Generation Model	Llama 3 8B Instruct	Llama 3 70B Instruct	Llama 3 8B Instruct (quant.)	Mistral 7B Instruct (quant.)
Parameter Count	8.03B	70.6B	8.03B	7.25B
Parameter Format	bf16	bf16	q4_0	q4_0
% Correct Format	83.34	83.33	92.95	99.86
Avg. question similarity	0.10	0.10	0.11	0.08
Avg. question length	51	56	51	60
Avg. answer length	722	601	804	667
HotPotQA accuracy	29%	30%	26%	11%
HotPotQA F1	0.42	0.39	0.36	0.26

Table 3. Synthetic dataset metrics and results on the PubMedQA benchmark.

Dataset Generation Model	Llama 3 8B Instruct	Llama 3 70B Instruct	Llama 3 8B Instruct (quant.)	Mistral 7B Instruct (quant.)
Parameter Count	8.03B	70.6B	8.03B	7.25B
Parameter Format	bf16	bf16	q4_0	q4_0
% Correct Format	66.16	66.67	74.17	99.90
Avg. question similarity	0.19	0.24	0.19	0.20
Avg. question length	80	83	78	88
Avg. answer length	1100	666	1039	1033
PubMedQA accuracy	60%	37%	50%	52%
PubMedQA F1	0.25	0.18	0.22	0.23

4.1 Model Factors

We used Llama 3 8B Instruct, a model with 8.03 billion parameters, in both its original bf16 and quantized q4-0 formats. We also used Llama 3 70B Instruct, of the same architecture, with 70.6 billion parameters in its original bf16 format. Finally, we used Mistral 7B Instruct, the smallest of the models with 7.25 billion parameters in quantized q4-0 format.

4.2 Percentage of Correctly Formatted Synthetic Samples

The unquantized models, Llama 3 8B Instruct and Llama 3 80B Instruct, generated questions and answers in the correct format at nearly identical rates: 83% for HotPotQA contexts and 66% to 67% for PubMedQA contexts. Llama 3 8B Instruct (quantized) generated in the correct format at higher rates: 93% for HotPotQA contexts and 74% for PubMedQA contexts. Mistral 7B Instruct (quantized) almost always generated in the correct format, above 99% for both datasets.

4.3 Average Question Similarity

All synthetic datasets had a normal distribution of cosine similarities between question pairs. Average question similarities were distinct between HotPotQA and PubMedQA datasets but were consistent within those groups. HotPotQA datasets ranged from 0.08-0.11 average cosine similarity, while PubMedQA datasets showed higher similarity between questions, ranging from 0.19 to 0.24 average cosine similarity. We also measured the percentage of question pairs within the same synthetic dataset that are very similar to each other (greater than 0.8 cosine similarity). In each dataset, this percentage was less than 0.1%, indicating that near-duplicate questions were rare.

4.4 Average Datapoint Length

Average of generated questions ranged from 51 to 60 characters for HotPotQA datasets and from 78 to 88 characters for PubMedQA datasets. Average length of generated

answers ranged from 601 to 804 characters for HotPotQA datasets and from 666 to 1100 characters for PubMedQA datasets.

4.5 Downstream Performance

On the HotPotQA benchmark, the datasets generated by Llama 3 models yielded similar results, ranging from 26% to 30% accuracy and 0.36 to 0.42 F1 scores. The dataset generated by the Mistral 7B Instruct (quantized) model achieved only 11% accuracy and a 0.26 F1 score downstream.

On the PubMedQA benchmark, accuracy rates varied from 37% to 60%, while F1 scores ranged from 0.18 to 0.25.

5 Discussion

In this section, we discuss the results and potential relationships between model factors, synthetic benchmark metrics, and downstream performance of fine-tuned models. We attempt to explain the results where possible and evaluate if our results demonstrate whether each metric would be useful in a synthetic dataset benchmark.

5.1 Model Factors

While our results do not show any definitive relationship between the size and parameter format (i.e., quantization) of the generation model with downstream performance, basic information such as these factors and the model architecture are important to note for a full understanding of a generated dataset. As described in [Hodak et al., 2023], benchmarks for large language models should include measures of accuracy, coherence, cost, energy efficiency, and throughput, among other factors. These model factors are important to know in the context of a synthetic dataset. A generative model's accuracy and coherence, especially graded from benchmarks similar in domain or task to the dataset being generated, can indicate whether the choice of model was appropriate to generate the dataset. For example, one should have low confidence in the quality of a synthetic dataset in the mathematical word problems domain that was generated by a model that scored poorly on the MMLU Math [Hendrycks et al., 2021] benchmark. Additionally, if changes or additions are needed in a synthetic dataset, it will be important to know the cost, energy requirements, and time requirements of further generation with the model. These factors are directly related to the size, format, and architecture of a language model.

5.2 Percentage of Correctly Formatted Synthetic Samples

Because incorrectly formatted synthetic samples were removed from the fine-tuning dataset, the percentage of correctly formatted samples is unlikely to have an impact on the downstream performance. Yet this metric is relevant for practical purposes, as models that follow the desired generation format at a low rate will require much more compute to generate the same amount of correctly formatted samples compared to a model

that dependably follows desired formats. However, this is an imperfect metric, as there is no guarantee that filtering efforts eliminate all format inconsistencies. While measuring this metric may not be necessary for synthetic datasets, the process of checking and cleaning synthetic datasets for the desired formatting and content is very important to the overall data quality and downstream performance. The unquantized models, Llama 3 8B Instruct and Llama 3 70B Instruct, underperformed the quantized models in matching the desired format. Mistral 7B Instruct (quantized) was the best model for generating data in the desired format. Although Mistral matched the desired format in over 99% of samples, its downstream performance on HotPotQA was the worst of any model, indicating that correct formatting of generated data does not guarantee that data demonstrates more complex qualities such as language understanding and knowledge synthesis.

5.3 Average Question Similarity

Average question similarity appears to depend on the dataset used as context (i.e., HotPotQA or PubMedQA). Within these groups, there was very little difference in similarity scores between datasets from different LLMs. Likely, domain-specific contexts, such as PubMedQA, will yield more similar questions than general-domain contexts like HotPotQA. This outcome is apparent in our results. The distribution of similarity scores between samples should be checked for synthetic datasets to ensure the dataset has the desired diversity, with the understanding that domain-specific datasets will likely have higher similarity scores. Otherwise, the downstream fine-tuned model may not learn to generalize to the task properly. While we found that near duplicates in our generated data were rare, it is important to check for these to ensure quality training data.

5.4 Average Datapoint Length

An unexpected result of our experiment was that the dataset generated by the Llama 3 70B Instruct model, the largest and most generally capable model used, based on PubMedQA context had the worst downstream performance on the PubMedQA benchmark. A salient metric for this dataset was its relatively short average answer length: at 666 characters, its answers were over 30% shorter than those of the other PubMedQA synthetic datasets. While we expected the largest model to generate the best synthetic dataset, in this case, the opposite occurred. We hypothesize that answer length is an indicator of quality when using chain-of-thought reasoning [Wei et al., 2023] in synthetic datasets. PubMedQA synthetic datasets showed a strong correlation between average answer length and downstream performance measures, with correlation coefficients 0.948 and 0.955 compared to accuracy and F1 scores, respectively. With a sample size of only four PubMedQA synthetic datasets, we believe short answer lengths are likely associated with lower quality training data, but this should be further studied.

Table 4 demonstrates an example question-and-answer pair from the PubMedQA synthetic dataset created with Llama 3 70B Instruct. This answer is the shortest in the dataset, and, although the final answer is correct, it does not demonstrate generalizable reasoning steps.

Table 4. "We need to look for the relevant information in the context," taken from the synthetic dataset with the worst downstream performance on the PubMedQA benchmark, does not demonstrate generalizable reasoning steps.

Prompt context (not generated)	From March 2007 to January 2011, 88 DBE procedures were performed on 66 patients. Indications included evaluation anemia/gastrointestinal bleed, small bowel IBD and dilation of strictures. Video-capsule endoscopy (VCE) was used prior to DBE in 43 of the 66 patients prior to DBE evaluation. The mean age was 62 years. Thirty-two patients were female, 15 were African-American; 44 antegrade and 44 retrograde DBEs were performed. The mean time per antegrade DBE was 107.4±30.0 minutes with a distance of 318.4±152.9 cm reached past the pylorus. The mean time per lower DBE was 100.7±27.3 minutes with 168.9±109.1 cm meters past the ileocecal valve reached. Endoscopic therapy in the form of electrocautery to ablate bleeding sources was performed in 20 patients (30.3%), biopsy in 17 patients (25.8%) and dilation of Crohn's-related small bowel strictures in 4 (6.1%). 43 VCEs with pathology noted were performed prior to DBE, with findings endoscopically confirmed in 32 cases (74.4%). In 3 cases the DBE showed findings not noted on VCE.
Generated question	How many patients were African-American?
Generated chain-of-thought answer	Here's the step-by-step reasoning to answer the question: To find the number of patients who were African-American, we need to look for the relevant information in the context. ##begin_quote## Fifteen were African-American; ##end_quote## This sentence directly provides the answer to the question. <ANSWER>: 15

We start with a context document from PubMedQA–this is an abstract from a medical research paper. The LLM is prompted with the context document, along with instructions to generate a relevant question and chain-of-thought answer. In this case, the LLM generated the question "How many patients were African-American?" The answer to this question can be found in the context document: the fifth sentence states 15 of the patients were African-American.

An answer that demonstrates strong reasoning ability, making it a good training example for model fine-tuning, would break down the task into intermediate steps relevant to the context of the question. For example: "To find the number of patients who were African-American, we need to look for patient demographics in the study. Then, we need to find the total number of African-American patients who underwent DBE procedures."

However, the answer generated by the LLM did not break the task into intermediate steps or frame thoughts in the context of the question. Instead, it simply states "To find

the number of patients who were African-American, we need to look for the relevant information in the context." This answer is very short, because it does not break down the task into components. As a training example, this answer does not effectively demonstrate chain-of-thought reasoning, likely leading to worse downstream performance of the fine-tuned LLM.

5.5 Model Recommendation

Our results show that average answer length of a synthetic question and answer dataset is the best indicator, of those considered, of downstream performance. Llama 3 8B Instruct (bf16) performed well in dataset generation for both open-domain (HotPotQA) and domain-specific (PubMedQA) contexts. In general, we recommend using this model for synthetic dataset generation.

5.6 Future Work

Future Research on Synthetic Dataset Benchmarking. Further work could be done to provide clear definitions for these and other metrics for synthetic dataset. A toolkit or set of best practices for cleaning synthetic datasets and ensuring they adhere to desired formats would likely improve their utility in downstream applications. Additionally, while we believe short chain-of-thought answer length may be an indicator of poor quality, this measurement is only a naïve proxy for the quality of a highly complex reasoning mechanism. While language models are being used to generate answers to questions and synthetic datasets, there is also likely great potential for language models to assist in measuring and improving the quality of natural language datasets.

Expanding the LLMs used in this study to other families of open-source models and to much larger (hundreds of billions of parameters) or smaller (1-2 billion parameters) models may corroborate or add nuances to our findings.

A Suggested Modification of RAFT. As a modification on RAFT as described in [Zhang et al., 2024], we propose that those implementing RAFT consider the specific task the synthetic dataset is supposed to exemplify. For example, the PubMedQA dataset's task is classification between "yes", "no", and "maybe" answers. However, the published RAFT code for generating synthetic datasets [Patil, 2024] yields freeform answers. While these synthetic datasets likely improve a fine-tuned LLM's in-domain reasoning, they do not train it to perform the task of classification. This disconnect between training and evaluation tasks is likely why the results of [Zhang et al., 2024] show that RAFT did not significantly outperform domain-specific fine-tuning + RAG, which, like their implementation of RAFT, would demonstrate in-domain reasoning, but not classification.

6 Conclusion

In this paper, we experimented with various language models to generate synthetic datasets within the RAFT process [Zhang et al., 2024]. Our findings enable a more

informed use of synthetic datasets, moving beyond the naïve acceptance of generative language model outputs. This work represents a significant step forward for RAFT and other processes relying on synthetic datasets, especially for fine-tuning large language models and other compute-intensive tasks.

Considering the rapid evolution of generative artificial intelligence (AI), it is crucial for researchers and industry experts to develop innovative techniques for evaluating AI-generated content. Our study on synthetic natural language dataset quality provides a foundation for comprehensive benchmarks. By offering a framework to assess the quality and limitations of these datasets, we aim to improve synthetic data quality and optimize the use of language model fine-tuning resources.

References

1. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of Age and Gender on Blogging. Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs (2006). https://u.cs.biu.ac.il/~schlerj/schler_springsymp06.pdf
2. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. arXiv:1506.06724 (2015). https://arxiv.org/abs/1506.06724
3. Zhang, T., Patil, S.G., Jain, N., Shen, S., Zaharia, M., Stoica, I., Gonzalez, J.E.: RAFT: Adapting Language Model to Domain Specific RAG. arXiv:2403.10131 (2024). https://arxiv.org/abs/2403.10131
4. Lin, X., Wang, W., Li, Y., Yang, S., Feng, F., Wei, Y., Chua, T.: Data-efficient Fine-tuning for LLM-based Recommendation. arXiv:2401.17197 (2024). https://arxiv.org/abs/2401.17197
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv:1706.03762 (2017). https://arxiv.org/abs/1706.03762
6. Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J.: Large Language Models: A Survey. arXiv:2402.06196 (2024). https://arxiv.org/abs/2402.06196
7. Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., Gadepally, V.: From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. arXiv:2310.03003 (2023). https://arxiv.org/abs/2310.03003
8. Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., Grave, E.: Atlas: Few-shot Learning with Retrieval Augmented Language Models. Journal of Machine Learning Research 24(251), 1-43 (2023). https://jmlr.org/papers/v24/23-0037.html
9. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 (2020). https://arxiv.org/abs/2005.11401
10. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H.: Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 (2023). https://arxiv.org/abs/2312.10997

11. Soudani, H., Kanoulas, E., Hasibi, F.: Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge. arXiv:2403.01432 (2024). https://arxiv.org/abs/2403.01432
12. Lin, X.V., Chen, X., Chen, M., Shi, W., Lomeli, M., James, R., Rodriguez, P., Kahn, J., Szilvasy, G., Lewis, M., Zettlemoyer, L., Yih, S.: RA-DIT: Retrieval-Augmented Dual Instruction Tuning. arXiv:2310.01352 (2023). https://arxiv.org/abs/2310.01352
13. Xu, P., Ping, W., Wu, X., McAfee, L., Zhu, C., Liu, Z., Subramanian, S., Bakhturina, E., Shoeybi, M., Catanzaro, B.: Retrieval meets Long Context Large Language Models. arXiv:2310.03025 (2024). https://arxiv.org/abs/2310.03025
14. Wang, B., Ping, W., McAfee, L., Xu, P., Li, B., Shoeybi, M., Catanzaro, B.: InstructRetro: Instruction Tuning post Retrieval-Augmented Pretraining. arXiv:2310.07713 (2023). https://arxiv.org/abs/2310.07713
15. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., Fidler, S.: DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort. arXiv:2401.10225 (2024). https://arxiv.org/abs/2401.10225
16. Li, Z., Zhu, H., Lu, Z., Yin, M.: Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. arXiv:2310.07849 (2023). https://arxiv.org/abs/2310.07849
17. Karras, T., Laine, S., Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv:1812.04948 (2018). https://arxiv.org/abs/1812.04948
18. Zhang, Y., Ling, H., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. arXiv:2112.10741 (2021). https://arxiv.org/abs/2112.10741
19. He, R., Sun, S., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Is Synthetic Data from Generative Models Ready for Image Recognition?. arXiv:2210.07574 (2022). https://arxiv.org/abs/2210.07574
20. Besnier, V., Jain, H., Bursuc, A., Cord, M., Pérez, P.: This Dataset Does Not Exist: Training Models from Generated Images. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE, Barcelona, Spain (2020). doi: 10.1109/ICASSP40776.2020.9053146
21. Wang, Z., Yu, A., Firat, O., Cao, Y.: Towards Zero-Label Language Learning. arXiv:2109.09193 (2021). https://arxiv.org/abs/2109.09193
22. Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T., Kong, L.: ZeroGen: Efficient Zero-shot Learning via Dataset Generation. arXiv:2202.07922 (2022). https://arxiv.org/abs/2202.07922
23. Kumar, V., Choudhary, A., Cho, E.: Data Augmentation using Pre-trained Transformer Models. arXiv:2003.02245 (2020). https://arxiv.org/abs/2003.02245
24. Chung, J.J.Y., Kamar, E., Rosenbaum, A., Kim, S., Ray, D., Amershi, S.: Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions. arXiv:2306.04140 (2023). https://arxiv.org/abs/2306.04140
25. Gupta, H., Scaria, K., Anantheswaran, U., Verma, S., Parmar, M., Sawant, S.A., Baral, C., Mishra, S.: TarGEN: Targeted Data Generation with Large Language Models. arXiv:2310.17876 (2023). https://arxiv.org/abs/2310.17876
26. Xu, J., Wu, H., Wang, J., Long, M.: Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. (2022). https://openreview.net/forum?id=aDC4benbIwL
27. Chen, M., Papangelis, A., Tao, C., Rosenbaum, A., Kim, S., Liu, Y., Yu, Z., Hakkani-Tur, D.: Weakly Supervised Data Augmentation Through Prompting for Dialogue Understanding. arXiv:2210.14169 (2022). https://arxiv.org/abs/2210.14169

28. Veselovsky, V., Ribeiro, M.H., Arora, A., Josifoski, M., Anderson, A., West, R.: Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science. arXiv:2305.15041 (2023). https://arxiv.org/abs/2305.15041
29. Guo, X., Chen, Y.: Generative AI for Synthetic Data Generation: Methods, Challenges and the Future. arXiv:2403.04190 (2024). https://arxiv.org/abs/2403.04190
30. Layeghy, S., Gallagher, M., Portmann, M.: Benchmarking the Benchmark – Analysis of Synthetic NIDS Datasets. arXiv:2104.09029 (2021). https://arxiv.org/abs/2104.09029
31. Singh, K., Navaratnam, T., Holmer, J., Schaub-Meyer, S., Roth, S.: Is Synthetic Data All We Need? Benchmarking the Robustness of Models Trained with Synthetic Images. arXiv:2405.20469 (2024). https://arxiv.org/abs/2405.20469
32. Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W., Mooney, S.D., Baral, C., Mishra, S.: A Multifaceted Benchmarking of Synthetic Electronic Health Record Generation Models. arXiv:2208.01230 (2022). https://arxiv.org/abs/2208.01230
33. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. arXiv:1809.09600 (2018). https://arxiv.org/abs/1809.09600
34. Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W., Lu, X.: PubMedQA: A Dataset for Biomedical Research Question Answering. arXiv:1909.06146 (2019). https://arxiv.org/abs/1909.06146
35. Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 (2021). https://arxiv.org/abs/2106.09685
36. Hodak, M., Ellison, D., Van Buren, C., Jiang, X., Dholakia, A.: Benchmarking Large Language Models: Opportunities and Challenges. In: TPC Technology Conference 2023, Vancouver, BC, Canada (2023)
37. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring Massive Multitask Language Understanding. arXiv:2009.03300 (2021). https://arxiv.org/abs/2009.03300
38. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 (2023). https://arxiv.org/abs/2201.11903
39. Patil, L.: RAFT [Source Code]. https://github.com/ShishirPatil/gorilla/tree/main/raft (2024)