Can LLMs Augment Low-Resource Reading Comprehension Datasets? Opportunities and Challenges

Vinay Samuel[1], Houda Aynaou[2], Arijit Ghosh Chowdhury[3], Karthik Venkat Ramanan[3], Aman Chadha[4†]

[1]Carnegie Mellon University, [2]Georgia Tech University,
[3]University of Illinois Urbana-Champaign, [4]Amazon GenAI
vsamuel@andrew.cmu.edu

Abstract

Large Language Models (LLMs) have demonstrated impressive zero-shot performance on a wide range of NLP tasks, demonstrating the ability to reason and apply common sense. A relevant application is to use them for creating high-quality synthetic datasets for downstream tasks. In this work, we probe whether GPT-4 can be used to augment existing extractive reading comprehension datasets. Automating data annotation processes has the potential to save large amounts of time, money, and effort that goes into manually labeling datasets. In this paper, we evaluate the performance of GPT-4 as a replacement for human annotators for low-resource reading comprehension tasks, by comparing performance after fine-tuning, and the cost associated with annotation. This work serves to be the first analysis of LLMs as synthetic data augmenters for QA systems, highlighting the unique opportunities and challenges. Additionally, we release augmented versions of low-resource datasets, that will allow the research community to create further benchmarks for evaluation of generated datasets. Github available at https://github.com/vsamuel2003/qa-gpt4

1 Introduction

Machine reading comprehension (MRC) is a challenging NLP task where systems are designed to answer questions based on a given context. This task has significant practical value, as it answers user queries in diverse settings, from clinical contexts (Krithara et al., 2023; Pampari et al., 2018; Pappas et al., 2020), to customer support (Castelli et al., 2020) and policy interpretation (Ahmad et al., 2020). BERT-based models (Glass et al., 2020) have achieved state-of-the-art performance when trained with extensive data from datasets like SQuAD (Rajpurkar et al., 2018) and Natural Questions (Kwiatkowski et al., 2019). However, their effectiveness diminishes in low-resource domains with limited data points (Schmidt et al., 2022). This limitation becomes particularly pronounced in newly emerging fields such as COVID-19 (Möller et al., 2020), where substantial annotated instances are often lacking.

Data augmentation has been instrumental in enhancing performance across numerous low-resource NLP tasks (Feng et al., 2021; Wang et al., 2022; Liu et al., 2021). Yet, much of the work on data augmentation for QA (Alberti et al., 2019; Shakeri et al., 2020; Bartolo et al., 2021; Dhingra et al., 2018; Yang et al., 2017), hinges on the availability of unlabeled paragraphs from common sources, such as Wikipedia, to produce new context-question-answer instances. This approach poses a challenge for specialized and mission-critical domains where such unlabeled contexts are scarcely available. Bridging this gap, Large Language Models (LLMs) exhibit a capability to generate texts that closely resemble human-authored content (Brown et al., 2020; Clark et al., 2021). This potential of LLMs can be harnessed to generate both novel contexts and their corresponding question-answer pairs.

Addressing this gap, we introduce a GPT-4 (OpenAI, 2023) based data augmentation technique tailored for low-resource machine reading comprehension, specifically focusing on the extractive setting. In extractive QA, the system is provided with a context passage and a question, and the system must determine if the question is answerable using an extractive span from the passage. Our approach begins by generating supplementary contexts, questions, and answers to augment training sets. To achieve this, we use in-context learning with passages, questions, and answers from the training set, ensuring minimal domain shift between the synthetically generated data and the original datasets.

Subsequently, we adopt cycle-consistent filtering to isolate high-quality training instances. Em-

[†]Work does not relate to position at Amazon.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 411–421

pirical evaluations conducted on three pertinent real-world low-resource datasets CovidQA (Möller et al., 2020), PolicyQA (Ahmad et al., 2020), and TechQA (Castelli et al., 2020) reveal that our methodology improves the performance of BERT-based MRC on CovidQA by 23% and on PolicyQA by 5% in terms of exact match. Notably, our approach attains state-of-the-art results on CovidQA.

2 Related Work

Language models have played a key role in the creation of synthetic datasets for various NLP tasks. Models such as GPT-2 (Radford et al., 2019) and CTRL (Keskar et al., 2019) have been applied to areas including general language understanding (Meng et al., 2022; He et al., 2022), classification (Kumar et al., 2020; Anaby-Tavor et al., 2019), dialogue tasks (Mohapatra et al., 2021), commonsense reasoning (Yang et al., 2020), and relation extraction (Papanikolaou and Pierleoni, 2020), among others. Recently, LLMs have significantly improved the quality and scope of synthetic dataset generation. They have been instrumental in augmenting datasets for tasks such as NLI and sentiment analysis (Dixit et al., 2022), classification (Yoo et al., 2021), and even creating datasets for personalized dialogue generation (Lee et al., 2022), hate speech detection (Hartvigsen et al., 2022), and textual similarity (Schick and Schütze, 2021) to name a few.

Most prior work in synthetic data generation for QA (Riabi et al., 2021; Chakravarti et al., 2020; Du and Cardie, 2018; Alberti et al., 2019) has concentrated on generating questions from Wikipedia passages to produce supplementary training examples. More recently, Kalpakchi and Boye introduced the use of GPT-3 for creating extra training data for Swedish multiple-choice questions. Our approach is the first to utilize in-context learning with LLMs for synthesizing contexts, questions, and answers for low-resource MRC.

3 Setup

3.1 Low Resource Datasets

We utilize three reading comprehension datasets in our work: CovidQA, PolicyQA, and TechQA. These datasets cover diverse domains while having relatively small training sizes, making them well-suited for evaluating synthetic data augmentation techniques.

The CovidQA dataset (Möller et al., 2020) focuses on question answering related to the COVID-19 pandemic. It contains 2,019 question-answer pairs on topics such as virus transmission, public health interventions, and social impacts.

PolicyQA (Ahmad et al., 2020) contains 12,102 question-answer pairs about United States immigration and travel policies. The questions require reasoning about specific policy documents to determine the answer.

TechQA (Castelli et al., 2020) provides 1,808 examples related to technical support issues on computer networking, software, and hardware. The goal is to develop QA systems that can resolve technical problems automatically.

In summary, these three datasets cover the domains of healthcare, public policy, and technology, while having relatively small training set sizes between 1-10k examples. This makes them suitable testbeds for studying the effects of augmenting the training data through synthetic example generation.

4 Synthetic Data Generation

We generate synthetic examples for each dataset using the in-context learning capabilities of the GPT-4 model. As part of our contribution, we release all synthetically augmented datasets to promote reproducibility and further research into refining the use of bootstrapping datasets with synthetically generated data. Dataset statistics are included in the Results section of this paper. Furthermore, examples of original data instances and synthetically generated data instances are included in the Appendix. The data generation process consists of two stages:

4.1 Context Generation

In the first stage, we provide GPT-4 with either 1 example (one-shot) or 2 examples (two-shot) of contexts from the original training set of each dataset. These few-shot examples prime GPT-4 on the style and topics present in the contexts. Providing just one or two examples allows GPT-4 to adapt from demonstrations due to the robust few-shot learning capabilities of LLMs (Reif et al., 2022; Frohberg and Binder, 2022; Wei et al., 2022). We then generate new synthetic paragraph-length contexts by providing a prompt and allowing GPT-4 to complete the paragraph based on the few-shot priming.

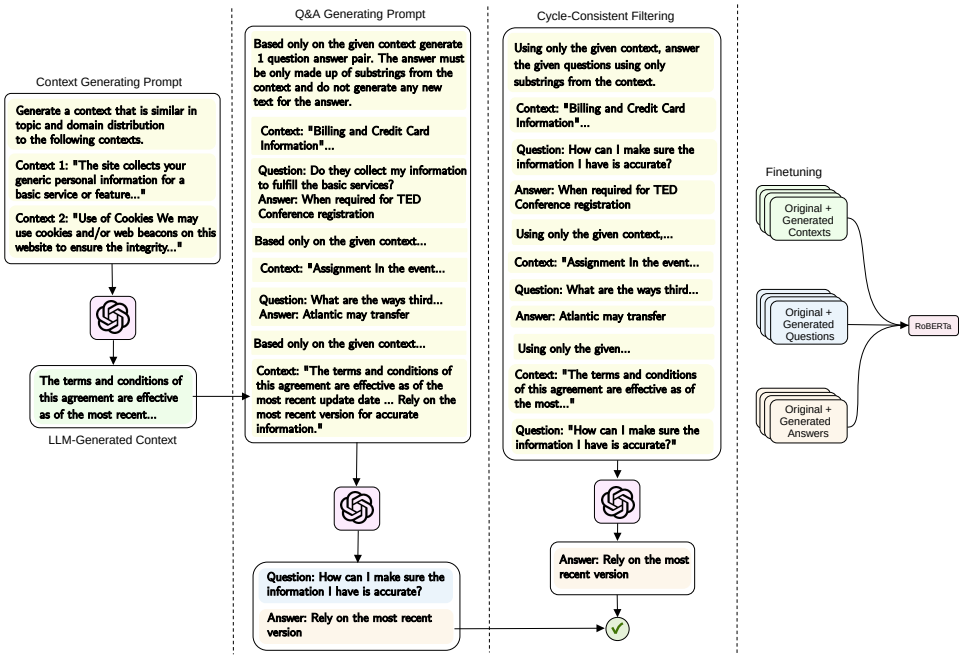

Figure 1: Overview of our methodology using PolicyQA as an example with 2-shot prompts.

4.2 QA Generation

The second stage generates synthetic question-answer pairs conditioned on the synthetic contexts. We again prime GPT-4 with either 1 example (one-shot) or 2 examples (two-shot) of QA pairs from the original dataset. The few-shot priming allows GPT-4 to learn the QA pattern quickly. We then provide the synthetic context from the first stage along with a prompt for GPT-4 to generate a relevant question-and-answer pair mimicking the style of the examples.

This two-stage process allows us to leverage the few-shot learning and text generation capabilities of GPT-4 to produce synthetic datasets that mimic the style and semantics of the original data. We generate varying amounts of synthetic data, from 1x to 10x the size of the original training sets, to study the impact on downstream task performance.

4.2.1 Round Trip Filtration

To further improve the quality of the synthetic QA pairs, we implement a round-trip filtration technique. After generating a synthetic question and answer using GPT-4, we provide the question back to the model without the answer. We allow GPT-4 to attempt to answer the question again based on the context. If the model's newly generated answer matches the original synthetic answer, we retain this QA pair, as it indicates a high-quality question with a consistent answer. If the answers do not match, we discard the synthetic QA pair under the assumption that the question is flawed in some way.

This round-trip filtration process provides a mechanism for GPT-4 to self-filter its own generated content. By only keeping QA pairs that exhibit consistency when answered twice, we obtain higher-quality synthetic data for downstream training. The filtration process improves precision at the potential expense of some recall.

4.3 Prompt Selection

We derived our final prompts for both the Context generation and the QA generation keeping certain design choices in mind. From preliminary experiments, it was noted that in the zero-shot setting, the GPT-4 model would generate contexts and QA pairs that were not from a similar distribution as the dataset to be augmented. This eventually led to downstream performance loss in the fine-tuning stage. To prevent this, n-shot prompting was used for in-context learning where $n = 1$ and $n = 2$ were experimented with. For the context generation phase, this meant prompting with n randomly selected contexts from the original datasets to generate the synthetic context, and for the QA generation this meant prompting the model with n randomly selected (context, question, answer) triplets from the original dataset along with the synthetically generated context.

CovidQA		
Setup	**Exact Match**	**F1 Score**
Original Trainset	25.81	50.91
Baseline	19.71	44.18
One Shot	30.82	57.87
Two Shot	31.18	55.64
One Shot (CC)	**31.90**	**58.66**
Two Shot (CC)	30.82	53.40

PolicyQA		
Setup	**Exact Match**	**F1 Score**
Original Trainset	30.56	58.15
Baseline	30.08	57.65
One Shot	**32.18**	**59.61**
Two Shot	30.97	59.12
One Shot (CC)	30.76	58.71
Two Shot (CC)	30.47	58.46

TechQA		
Setup	**Exact Match**	**F1 Score**
Original Trainset	11.11	39.45
Baseline	**44.44**	**59.92**
One Shot	22.22	36.91
Two Shot	11.11	36.50
One Shot (CC)	22.22	41.76
Two Shot (CC)	22.22	44.73

Table 1: Experimental results for MRC across various datasets and settings.

4.4 Experiments

We train an extractive reading comprehension model using RoBERTa-Base with a learning rate of $3e-5$, batch size of 16, for 5 epochs. The model is implemented with Hugging Face and runs on an Nvidia V100 GPU, measuring F1 and Exact Match scores. For the baseline, we use a T5-based question generation model trained on the SQuAD dataset, which generates question-answer pairs from a paragraph.

5 Results

Table 1 presents results across three datasets. For the CovidQA dataset, we saw steady improvements in question-answering performance by augmenting the training set with synthetic data generated by GPT-4. The original training set achieved baseline exact match (EM) and F1 scores. Adding one-shot synthetic examples improved both metrics, with further gains observed using two-shot synthetic data. The highest EM and F1 scores were obtained with one-shot synthetic data combined with round-trip filtration, significantly surpassing the original training set.

For PolicyQA, the largest dataset with over 12,000 examples, the best performance was achieved by augmenting with one-shot synthetic data without filtration, improving EM by 1.6 points and F1 by 1.5 points over the baseline. This approach outperformed both two-shot and cycle-filtered variations.

In the smallest dataset, TechQA, with only 1,808 examples, synthetic data augmentation did not lead to clear improvements. The baseline model achieved the highest EM score, with two-shot cycle filtered, one-shot filtered, and one-shot unfiltered configurations performing similarly. For F1, two-shot cycle filtered data obtained the second-highest score after the baseline.

Overall, synthetic data augmentation improved performance in CovidQA and PolicyQA, with the best results from one-shot generation combined with round trip filtration for CovidQA, and unfiltered one-shot generation for PolicyQA. In TechQA, the small data size and high domain diversity limited the effectiveness of synthetic augmentation

Dataset statistics for the three datasets used are shown in Table 2 located in the appendix.

6 Opportunities and Challenges

Our experiments demonstrate the significant potential of leveraging LLMs like GPT-4 for synthetic data generation. In domains like CovidQA and PolicyQA, augmenting with LLM-generated synthetic examples consistently improved performance over the baseline, showcasing the few-shot generalization abilities of modern LLMs. One-shot synthetic data augmentation yielded the best results, surpassing other configurations. LLMs can significantly expand limited training sets for various NLP tasks, enhancing performance without the expense of human labeling.

However, challenges remain, particularly in low-data regimes like TechQA, where LLM-augmented models performed no better than the baseline. This highlights the difficulty LLMs face in synthesizing useful examples from scarce data. Improving LLMs' few-shot learning, integrating external knowledge, and developing advanced filtering techniques are critical for maximizing the benefits of synthetic data generation. While LLMs hold promise for addressing limited training data, substantial challenges must be overcome to fully realize their potential in diverse NLP tasks.

References

Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. PolicyQA: A reading comprehension dataset for privacy policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 743–749, Online. Association for Computational Linguistics.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, N. Tepper, and Naama Zwerdling. 2019. Do not have enough data? deep learning to the rescue! In *AAAI Conference on Artificial Intelligence*.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, Scott McCarley, Michael McCawley, Mohamed Nasr, Lin Pan, Cezar Pendus, John Pitrelli, Saurabh Pujar, Salim Roukos, Andrzej Sakrajda, Avi Sil, Rosario Uceda-Sosa, Todd Ward, and Rong Zhang. 2020. The TechQA dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1269–1278, Online. Association for Computational Linguistics.

Rishav Chakravarti, Anthony Ferritto, Bhavani Iyer, Lin Pan, Radu Florian, Salim Roukos, and Avi Sil. 2020. Towards building a robust industry-scale question answering system. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 90–101, Online. International Committee on Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 582–587, New Orleans, Louisiana. Association for Computational Linguistics.

Tanay Dixit, Bhargavi Paranjape, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. CORE: A retrieve-then-edit framework for counterfactual data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2964–2984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917, Melbourne, Australia. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Jörg Frohberg and Frank Binder. 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140, Marseille, France. European Language Resources Association.

Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, G P Shrivatsa Bhargav, Dinesh Garg, and Avi Sil. 2020. Span selection pre-training for question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2782, Online. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for implicit and adversarial hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. Generate, annotate, and learn: NLP with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842.

Dmytro Kalpakchi and Johan Boye. 2023. Quasi: a synthetic question-answering dataset in Swedish using GPT-3 and zero-shot learning. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 477–491, Tórshavn, Faroe Islands. University of Tartu Library.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *ArXiv*, abs/1909.05858.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. PERSONACHATGEN: Generating personalized dialogues using GPT-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*.

Biswesh Mohapatra, Gaurav Pandey, Danish Contractor, and Sachindra Joshi. 2021. Simulated chats for building dialog systems: Learning to generate conversations from instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1190–1203, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Yannis Papanikolaou and Andrea Pierleoni. 2020. Dare: Data augmented relation extraction with gpt-2. *ArXiv*, abs/2004.13845.

Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic data augmentation for zero-shot cross-lingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7016–7030, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maximilian Schmidt, A. Bartezzaghi, Jasmina Bogojeska, Adelmo Cristiano Innocenza Malossi, and Thang Vu. 2022. Improving low-resource question answering using active learning in multiple stages. *ArXiv*, abs/2211.14880.

Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460, Online. Association for Computational Linguistics.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022. PromDA: Prompt-based data augmentation for low-resource NLU tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4242–4255, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.

Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. 2017. Semi-supervised QA with generative domain-adaptive nets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1040–1050, Vancouver, Canada. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Dataset Statistics

Dataset	Original	One Shot	Two Shot	One Shot CC	Two Shot CC	Baseline
TechQA	388	775	775	775	775	768
PolicyQA	17056	68130	60306	63704	63704	51267
CovidQA	1461	6699	6716	6316	6316	8069

Table 2: Statistics for the dataset sizes of fine-tuning data for each experimental setting. The original category describes the original training data for each dataset before synthetic augmentation. One Shot CC and Two Shot CC show the round trip filtration applied on the One Shot and Two Shot augmented datasets.

B Qualitative Examples

Qualitative Examples from CovidQA

Original

Question: Why might we underestimate the spread of COVID19?

Answer: limited information available regarding incubation time, transmissibility, and virus origin

Context: ...
Abstract: 2019 Novel Coronavirus (2019-nCoV) is an emerging infectious disease closely related to MERS-CoV and SARS-CoV that was first reported in Wuhan City, Hubei Province, China in December 2019. As of January 2020, cases of 2019-nCoV are continuing to be reported in other Eastern Asian...
2 These numbers are likely underestimates due to the limited information available regarding incubation time, transmissibility, and virus origin. The What was the research question? Investigators adapted the "Identify, Isolate, Inform" (3I) Tool for use in suspected cases of 2019-nCoV....

One shot

Question: What is a significant concern in many parts of the Middle East regarding the consumption of meat and milk from Domestic Camels?

Answer: the potential for the transmission of MERS-CoV, a virus found in DC.

Context: The consumption of meat and milk from Domestic Camels (DC) is common in many parts of the Middle East, which has led to a significant concern regarding the potential for the transmission of MERS-CoV, a virus found in DC. However, the occurrence of MERS-CoV infection in the general public is reportedly much lower than expected given the widespread consumption of DC products....

Two shot

Question: What have research studies around 2016 and 2017 discovered about Ebola and Zika virus?

Answer: instances of genetic recombination within these virus strains.

Context: Several research studies in 2016 and 2017, involving the study of Ebola virus variants from West Africa and a variant of Zika virus from Brazil, have discovered instances of genetic recombination within these virus strains...

One shot CC

Question: What modern tools enable the collection of data in real-time for identifying disease outbreaks?

Answers: Modern digital tools

Context: Modern digital tools now enable the collection of data in real-time, significantly fast compared to traditional epidemiological methods that could take months to pinpoint the kind of outbreak [52]....

Two shot CC

Question: What have recent studies on Zika virus focused on?

Answers: exploring variants of the Zika virus have raised concerns about genetic recombination in viral distributions.

Context: Recent studies conducted in 2018 exploring variants of the Zika virus have raised concerns about genetic recombination in viral distributions. With human and mosquito genomes sharing over 99% identity, gene exchange can occur under certain biological conditions and co-infections, potentially leading to the emergence of unknown diseases [175] [185]....

Qualitative Examples from PolicyQA

Original

Question: Does the company collect any data that is personally identifiable?

Answer: personally identifiable

Context: Privacy Policy Posted: Tuesday, September 18, 2012 12:29 pm Updated: Thursday, May 21, 2015 3:22 pm. Our commitment to you The Frederick News-Post and its owner, Randall Family, LLC (collectively, "The News-Post"), recognize the importance of protecting the privacy of personally identifiable information collected about you, our customers. This customer privacy policy is intended to outline our legitimate business interests in collecting and using information and your reasonable expectations of privacy.

One shot

Question: What is Animoto's stance on privacy?

Answer: At Animoto, we prioritize your right to maintain control over your personal data.

Context: At Animoto, we prioritize your right to maintain control over your personal data. Our Privacy Policy, which is detailed below, outlines the steps taken by Animoto to ensure your privacy is safeguarded while you use the Animoto platform. We do like to remind you that the Animoto platform allows users who are 13 years or older to upload personal content (such as photos or information about themselves or others for whom they have obtained authority or consent to share such data) that could potentially be viewed and commented on by the public in forums or through private communication.

Two shot

Question: What does SnapDoodle offer to Registered Users?

Answer: SnapDoodle offers the capability for Registered Users and Subscribers to establish individual personal portfolios

Context: SnapDoodle offers the capability for Registered Users and Subscribers to establish individual personal portfolios either directly through SnapDoodle or via alternative sign-in facilities (e.g., Google Sign-In). In an effort to enhance user engagement, we ask for and show certain personal details to others in order to enable our users to recognize each other. Registered Users and Subscribers have the option to modify their portfolio information at any moment and can govern the way the service engages with them.

One shot CC

Question: What is the primary concern of Animagic?

Answers: your privacy

Context: At Animagic, your privacy is our primary concern. We acknowledge the importance of protecting your personal information and respect your rights to maintain control over its usage. The Privacy Policy outlined below is designed by Animagic to ensure your privacy is safeguarded while you navigate through Animagic Sites. It's worth noting that Animagic Sites and Services permit users aged 13 or above to upload personal content such as photos and information about themselves and others (provided they have requisite authority or approval). This content can be accessible for public scrutiny and discussion, as well as for private communication.

Two shot CC

Question: How can registered members and users create profiles on Slack?

Answers: either directly through Slack or through third-party log-in services

Context: Slack provides an option for its registered members and users to create personalized profiles either directly through Slack or through third-party log-in services like Google Connect. Aiming to enhance user interaction, we collect and display certain personal data to facilitate user identification among each other. Registered members or users have the flexibility to modify their profile details at any given time and have full control over their communication preferences with the service.

Qualitative Examples from TechQA

Original

Question: I am working on datacap 8.1 version. I have 2 pages in a batch. Page 1 with status as 1, Page 2 with status as 20. Both are of same page type. In verification step (dotedit) I want to hide the page which is with status 20. I have given status 20 in ignored page status in web administrator for verify. But still when I navigate through pages in dotedit batch view the respective panel is loading. I am able to submit the batch even if I have one page in a batch with 20 as status.there any way to hide the page (we have multiple same type of pages in a batch) from dotedit based on the page status?

Answer: The DeleteChildType from the Validations library can be used by running it from the parent of the object to be deleted and specifying the specific child. example, to remove blank pages from a batch: 1. Create a special page type attached at the batch level called something like DeleteMe. 2. Use an action like BlankPagesIDBySize to identify the blank page: * BlankPagesIDBySize (1000,DeleteMe)3. Create a rule with the action DeleteChild-Type(DeleteMe) and attach it at the batch level to have it remove all pages with page type DeleteMe.building a custom action, the DeleteChild API method is invoked from the parent object.

Context: dco document hierarchy node delete remove hide blank page TECHNOTE (FAQ)How do I delete a document hierarchy node, such as a blank page, so that it no longer processes rules or appears in a batchview listing? is sometimes desired to remove pages or documents from a batch, as they are no longer needed or to simplify processing for a Verify operator.DeleteChildType from the Validations library can be used by running it from the parent of the object to be deleted and specifying the specific child. example, to remove blank pages from a batch: 1. Create a special page type attached at the batch level called something like DeleteMe. 2. Use an action like BlankPagesIDBySize to identify the blank page: * BlankPagesIDBySize (1000,DeleteMe)3. Create a rule with the action DeleteChildType(DeleteMe) and attach it at the batch level to have it remove all pages with page type DeleteMe.building a custom action, the DeleteChild API method is invoked from the parent object. * *

One shot

Question: the ITCAM MQ Monitoring agent, we have a situation that generates alerts when a 2085 event (object unknown) occurs. We have recently seen alerts for the queue SYSTEM.MQXR.COMMAND.QUEUEfound following technote:Object Name [2085], SYSTEM.MQXR.COMMAND.QUEUE://www-01.ibm.com/support/docview.wss?uid=swg21681687technote does not mention Tivoli monitoring product, and only mentions monitoring products such as Nastel and InfraRed360.Tivoli monitoring agent for WebSphere MQ use the SYSTEM.MQXR.COMMAND.QUEUE? We are try to find out which application is causing the 2085 event.

Answer: Use the runmqsc display connection command to find the process id (PID) and application name. the above example of the queue Q1, this is the complete command to invoke under runmqsc: conn(*) where(objname eq Q1) all

Context: Identify application program connected queue TECHNOTE (TROUBLESHOOTING)(ABSTRACT)Your WebSphere MQ queue manager will not stop if there are applications that still have a queue opened. Your goal is to allow a graceful stop of the queue manager, also called controlled (or quiesced) shutdown...the runmqsc display connection command to find the process id (PID) and application name. the above example of the queue Q1, this is the complete command to invoke under runmqsc: conn(*) where(objname eq Q1) alloutput:8276: Display ...

Two shot

Question: Can I apply a TIP 2.2 fix pack directly to a TIP 2.1 installation?

Answer: In order to apply TIP 2.2 fix packs, the target TIP installation must already be at TIPCore 2.2.0 or newer. TIP 2.1 installations must be upgraded to TIP 2.2 using the TIP 2.2.0.1 feature pack.

Context: TIPL2; TIPL2INST; tivoli Integrated portal; feature pack TECHNOTE (FAQ)Can Tivoli Integrated Portal 2.2 fix packs be applied directly to a TIP 2.1 installation?order to apply TIP 2.2 fix packs, the target TIP installation must already be at TIPCore 2.2.0 or newer. TIP 2.1 installations must be upgraded to TIP 2.2 using the TIP 2.2.0.1 feature pack. The TIP 2.2.0.1 feature pack can be acquired from IBM Fix Central....

C Prompts

Prompts

Context Generation
"Generate a context that is similar in topic and domain distribution to the following contexts: {context1}, {context2}"

QA Generation
"Generate 1 question-answer pair. The answer must be only made up of substrings from the context and do not generate any new text for the answer. {n-shot context, question, answer triplets} Context:"