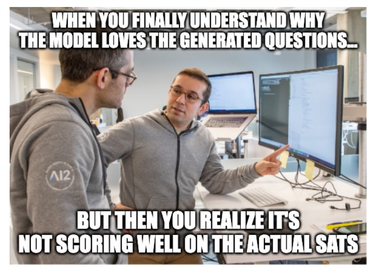
Question Mastery Unleashed: Self-Generated Data Augmentation is All You Need for AI to Ace Exams

Aydan Huang, Tian Zhou, Viola Xu



Background

❖ Intuition

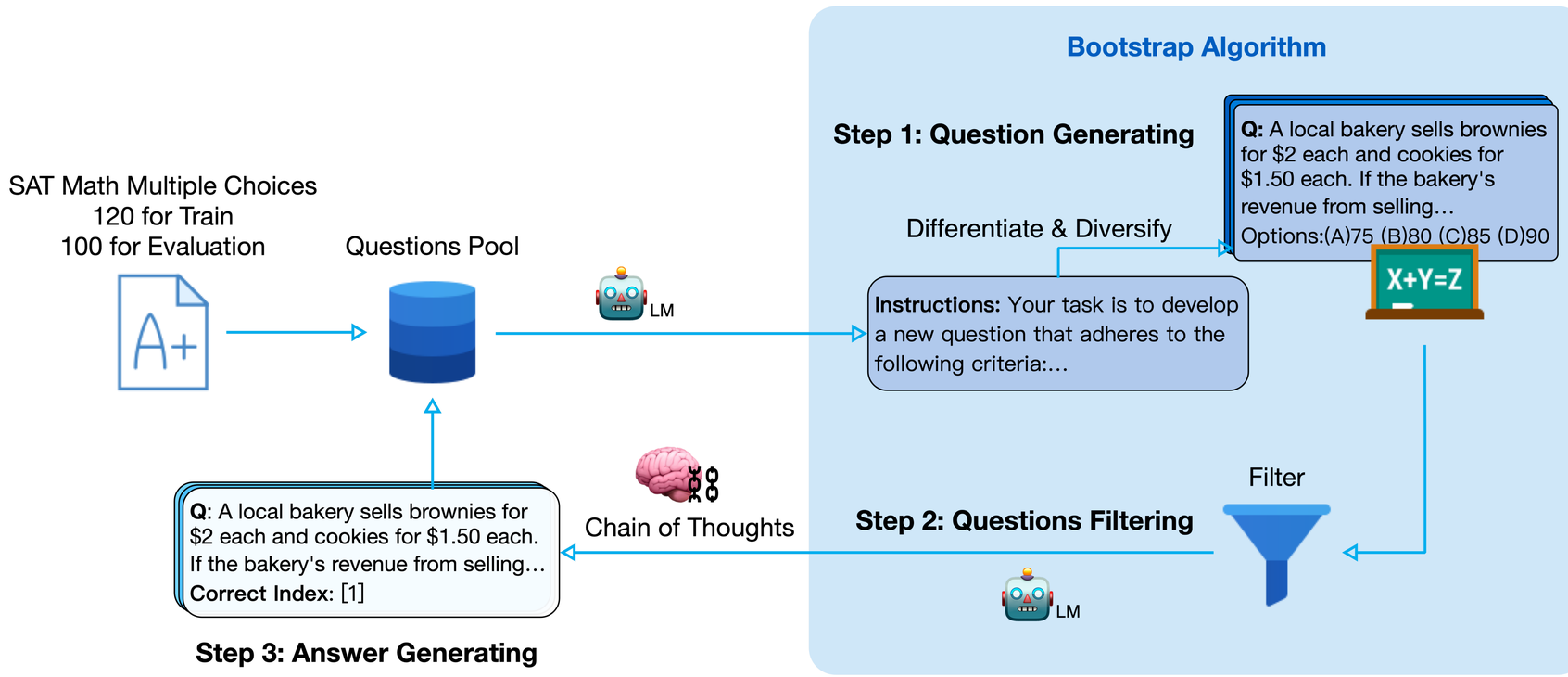
- Exam Datasets like SAT are restricted and impossible to obtain.
- Good Datasets are rare and expensive
- Find a way of natural language augmentation to enhance model performance

❖ Previous Work:

- Self-Instruct: introduces a framework to improve instruction-tuning's capabilities in large language models by generating new instruction data from the task seeds.
- Chain-of-Thoughts: Enhance accuracy and consistency of question answer generating.

❖ Idea: Instance Generation Pipeline!

Experiment



❖ Dataset

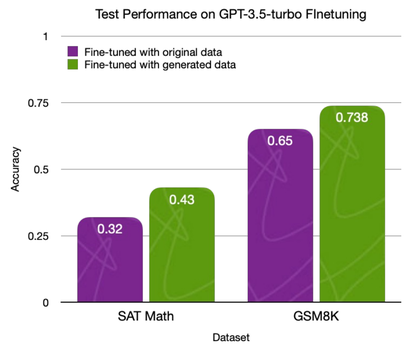
Attribute	SAT Dataset	GSM8K Dataset	AquaRat Dataset
Size of question seed	120	120	-
Size of generated dataset	1,000	2,000	-
Size of test set	100	1,318	234
Domain	High School Maths	Grade School Maths	Grad School Maths
Question Type	Multiple Choice	Short Answer	Multiple Choice
Source of Questions	SAT Exams	Crowdsourced	GMAT & GRE
Question Complexity	High	Elementary	Very High
Additional Features	Answer Options	Contextual Information	-

- We test this process on two models and two datasets.
- Models: Vanilla GPT3.5, Bert-Small
- Datasets: SAT Math, GSM8K

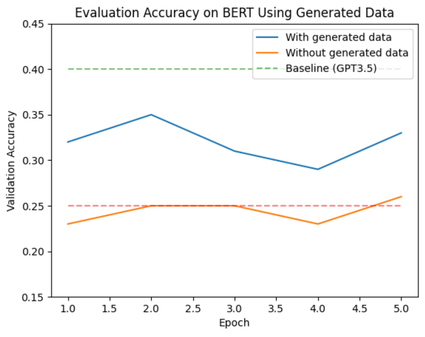
Result

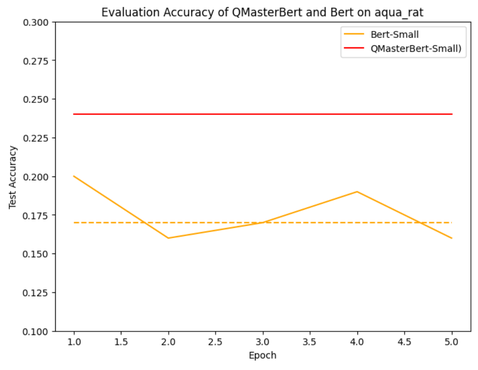
❖ Finetuned QMaster Model

- We Finetuned Bert and GPT3.5 with 1000 instances of self-generated SAT dataset and 1000 instances of self-generated GSM8K dataset
- QMasterGPT3.5
  - Better performance than the GPT3.5 tuned with original data on SAT Math and GSM8K
- QMasterBert-Small
  - Better performance than Bert-Small tuned with original data on SAT Math Multiple Choices
  - test on Aqua_rat, another multiple choices dataset does NOT train/tune before. QMasterBert$_{SAT}$ performs better than Bert-Small.



Model	Dataset	Accuracy
Bert-Small	SAT	0.26
QMasterBert-Small	SAT	0.33
Bert-Small	Aqua-rat	0.18
QMasterBert-Small	Aqua-rat	0.24
GPT3.5	SAT	0.32
QMasterGPT3.5	SAT	0.43
GPT3.5	GSM8K	0.65
QMasterGPT3.5	GSM8K	0.74





References:
[1] Wang et al., "Self-Instruct: Aligning Language Models with Self-Generated Instructions." https://arxiv.org/pdf/2212.10560.pdf
[2] YuLan-Chat-Team. 2023. Yulan-chat: An open- source bilingual chatbot. https://github. com/RUC-GSAI/YuLan-Chat.
[3] Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., \& Lim, E.-P. (2023). Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. arXiv preprint arXiv:2305.04091. Retrieved from https://arxiv.org/abs/2305.04091

Model

❖ Instruction

- Distinct Instructions for question generating, filtering and answer generating
- Instructions were carefully written, indicating the expected outcome and format.
- During the answer generating process, we use Chain of Thoughts Method.
- Multiple iterations of editing and review were done via LMs and human evaluation.

Instruction for Question Generating

Your task is to develop a new question that adheres to the following criteria:
1. Content and Style: …
2. To generate, inspired by the given question but … [diversify & differentiate] …
3. The question must be clear, logical, and answerable.

❖ Generated Questions and Answers

Given Q&A

Q: A rectangle was altered by **increasing its length by 10 percent** and decreasing its width by p percent. If these alterations decreased the area of the rectangle by 12 percent, what is the value of p ?
Answer Choices: (A)12 (B)15 (C)20 (D)22"

Generated Q&A

Q: A rectangular yard **is 3 times as long as it is wide.** If the perimeter of the yard is 64 meters, what is the area of the yard?
Answer Choices: (A) 108 square meters (B) 144 square meters (C) 192 square meters (D) 324 square meters"

Given Q&A

Q: A triangle has **angles of 30°, 60°, and 90°. If the length of the hypotenuse is 10 units**, what is the area of the triangle?
Answer Choices: (A) 25√3 square units (B) 12.5√3 square units (C) 50√3 square units (D) 100 square units

Generated Q&A

Q: In triangle PQR, **angle P measures 30 degrees, angle Q measures 75 degrees, and angle R measures 45 degrees**. If the length of side PQ is 10 units, what is the length of side QR?
Answer Choices: \n(A) 5√2 units \n(B) 10 units \n(C) 10√2 units \n(D) 15 units"

- The results demonstrate that the process is effective across various styles of questions, including multiple-choice and short-answer formats. The process is generalizable to different models, such as GPT-3.5 and BERT. We observe improved performance in these models across various datasets after fine-tuning with the generated dataset, indicating that the data produced by the pipeline is both usable and meaningful.