Assignment: Synthetic Data Generation for AI Systems

One often faces the challenge of limited datasets being available to train the model for any task in hand, be it Intent detection, slot filling, recommendation algorithm etc. The goal of this assignment is to generate synthetic datasets that can be used for training, designing, and evaluating such AI systems or tasks. The generated data should mimic realistic human-written text while allowing for controlled variations in factors such as dialogue length, topic diversity, and language complexity (This is not an exhaustive list, one can come up with new ones depending on the problem statement in hand).

Dataset

Below is the link to a subset of an Amazon reviews dataset, which comprises of reviews written by buyers on different products listed on amazon. To simplify the problem statement. I have only shared reviews for the following product category – "Supplements/Vitamins".

Link to a subset of the dataset (relevant to the problem) - Link to dataset (subset)

Link to the entire open-source dataset - https://amazon-reviews-2023.github.io/main.html

Tasks

Your job is to implement a methodology to generate a synthetic dataset of such reviews and along with that write answers to a couple of questions.

1. Why was the model/architecture used?
2. What were the different factors considered for generating this dataset? (Length, topic diversity etc.)
3. How do we measure the efficacy of a synthetic dataset?
4. How do we ensure the synthetic dataset one generates is inspired from a source dataset but not an exact replica?
5. What were the top challenges in solving for this problem statement?

Output

1. Repository Link
2. Generated dataset – excel/csv file with generated examples.
3. Written report / documentation file – A writeup of the steps followed – (any data cleansing work), experiments done, results documented, any specific insights. One can even add the different factors considered while generating dataset. This should also include the answers to the questions mentioned above.

Evaluation Criteria

1. Methodology – What all steps were followed? Which models/architectures were experimented with? Reasoning behind using the following architectures? How extensive was the research? etc.
2. Documentation

Note

This is a test of one's ability to research, experiment, and publish. Even if the results don't look good, it is important knowing where/how it could be improved. Feel free to use additional datasets/resources (just don't forget to mention that into the report).