# Assignment: Synthetic Data Generation for AI Systems

**By Jay Saraf**

1. There are two files given ‘product_asin.csv’ and 'reviews_supplements.csv'. I first renamed the title in both the files by changing the title header for the review to review_title and to the product to product_title.
2. Then I tried merging both the datasets on the basis of parent_asin and simultaneously converted the combined dataset into pandas data framework.
3. As only ‘Vitamins & Supplements’ subcategory was to be seen so I removed all the main_categories (according to the original Amazon Dataset 2023) and the sub categories which didn’t match this.
4. Further I checked whether there are some NaN values in any columns. I didn’t find any NaN values.

Architecture

1. Llama 3.2 ‘s 1B instruct and 3b instruct models were used in this assignment. I used these models because these are light weight which means they require less computation and they might resemble (not exactly, but close) to results of bigger models. I used Kaggle’s GPU P100 for 3b instruct and GPU T4 x 2 for 1b instruct.
2. First for the 1 B model I tried generating all the features(attributes) simultaneously by giving a single prompt. But the results weren’t that good and most of the time the synthetic data generated wasn’t unique. The length of the expected text for each feature was considered by approximating the size of the original data.
3. Rouge, BLEU score and cosine similarity could be used for finding out efficacy of a synthetic dataset.
4. I tried giving the adequate prompt such that the model takes inspiration from the given dataset but does not copy the exact data.
5. One of the challenges was removing the irrelevant data from the categories. It was important figure out a way to make the attributes related to each other.
6. If time was more I could have tried other models like Mistral, Gemma,etc. Further fine tuning using methods like lora could have been done.

References

1. https://huggingface.co/collections/meta-llama/llama-32-66f448ffc8c32f949b04c8cf
2. https://arxiv.org/html/2406.13130v1
3. https://stackoverflow.com/questions/69609401/suppress-huggingface-logging-warning-setting-pad-token-id-to-eos-token-id
4. https://arxiv.org/html/2404.07503v1