# **Synthetic Data generation from pandas Dataframe with LLM**

### **Model used : [nemotron-4-340b-instruct](https://build.nvidia.com/nvidia/nemotron-4-340b-instruct)**

### **Synthetic data generation :** unsuccessful

### **library used :** langchain, openai

### **API used :** nvidia opensource API because openai and Vertex AI are paid

I designed whole architecture of input prompt template data preprocessing and cleaning also but after using synthetic\_data\_generator of langchain I got suspicious Inference error Which I am trying now... to get Synthetic Data.

#### **Step 1 Amazon product review dataset cleaning and Converting into dictionary**

df = df.head(20)

# Converting the DataFrame to a list of dictionaries

examples = df.to\_dict(orient='records')

examples[0]

{'title\_x': 'Bariatric Fusion Bariatric Multivitamin Soft Chew | Tropical Fruit Flavor | Chewy for Post Bariatric Surgery Patients Including Gastric Bypass and Sleeve Gastrectomy | 60 Count | 1 Month Supply',

'parent\_asin': 'B0BKC2WYWB',

'categories': 'Health & Household',

'cat1': ' Heart Health Event',

'cat2': nan,

'cat3': nan,

'cat4': nan,

'cat5': nan,

'rating': 5,

'title\_y': 'Good vitamins',

'text': 'Is a bariatric patient, my vitamins are really important!! These Taste pretty good and seems to work.',

'asin': 'B07JH63HWS',

'user\_id': 'AF2HY3SCRK45T2ATV7FKRYUJCCTA',

'helpful\_vote': 0,

'verified\_purchase': True,

'date': Timestamp('2019-12-12 00:00:00'),

'time': datetime.time(12, 43)}

#### **Step 2 Defining Prompt template**

# Template for the synthetic data example

SYNTHETIC\_DATA\_EXAMPLE = PromptTemplate(

input\_variables=["title\_x", "parent\_asin", "categories", "cat1", "cat2", "cat3", "cat4", "cat5", "rating", "title\_y", "text", "asin", "user\_id", "helpful\_vote", "verified\_purchase", "date", "time"],

template=template\_str

)

prompt\_template = FewShotPromptTemplate(

prefix=SYNTHETIC\_FEW\_SHOT\_PREFIX,

examples=examples,

suffix=SYNTHETIC\_FEW\_SHOT\_SUFFIX,

input\_variables=["subject","extra"],

example\_prompt=SYNTHETIC\_DATA\_EXAMPLE,

)

#### **Step 4 Defining synthetic\_data\_generator**

synthetic\_data\_generator = create\_openai\_data\_generator(

output\_schema=AmazonProductReviews,

llm=ChatNVIDIA(model="nvidia/nemotron-4-340b-instruct",temperature=0.2),

prompt=prompt\_template

)

#### **Step 5 Trying to generate result**

synthetic\_results = synthetic\_data\_generator.generate(

subject="amazon\_product\_reviews",

extra = "the name must be chosen at random. Make it something you wouldn't normally choose.",

runs=10,

)

## **Why this Model ?**

* Nvidia's nemotron-4-340b-instruct LLM model is specially designed for synthetic data generation.It is a fine-tuned version of the Nemotron-4-340B-Base model, optimized for English-based single and multi-turn chat use-cases. It supports a context length of 4,096 tokens.
* The base model was pre-trained on a corpus of 9 trillion tokens consisting of a diverse assortment of English based texts, 50+ natural languages, and 40+ coding languages.

## **Efficiency measures and testing**

* We can measure efficiency of dataset by testing synthetic data as testing data against input data as train and validation data and check confusion Matrix for that.

## **Challenges**

* Lack of proper documentations
* almost all LLM services are paid
* Synthetic Data generation as csv is need to be developed

## **References**

[nemotron-4-340b-instruct LLM Model](https://build.nvidia.com/nvidia/nemotron-4-340b-instruct)

[Synthetic Data generation YouTube](https://www.youtube.com/watch?v=hMjtdECXlYo)

[Langchain Documentation on ChatNVIDIA](https://python.langchain.com/docs/integrations/chat/nvidia_ai_endpoints/)

[Amazon Product Reviews Dataset](https://drive.google.com/file/d/19eTFRj2ctWYOmdYHuC7h7qlBBDYqSVVM/view)