Report on Synthetic Dataset Generation for Reviews

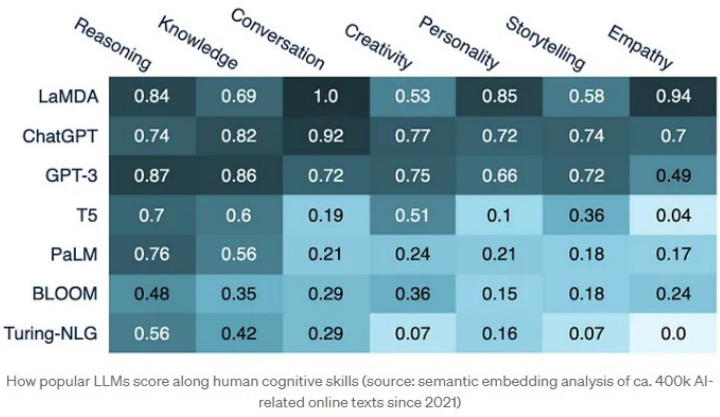
Sagar, Kumar

06-10-2024

Introduction In this report, we outline the methodology used to generate a synthetic dataset of reviews, detailing the model and architecture chosen, the factors considered during dataset generation, and the evaluation metrics for measuring efficacy. We also discuss strategies for ensuring originality and highlight the challenges faced during the process.

Methodology for Synthetic Dataset Generation The process of generating synthetic datasets can be broken down into several key steps:

1. **Data Collection:** Gathered a substantial amount of real reviews (Downloaded from the source url attached with the assignment instructions (GDrive)).
2. **Model Selection:** Choose a generative model suited for text generation. We opted for a transformer-based architecture, GPT-4o (Generative Pre-trained Transformer), due to its proficiency in understanding and generating human-like text. We had a comparision of LLMs to mimic humans in writing unique and diverse text. Here's a reference comparision table (which we referred to identify most suitable LLM) which prompted us to use GPT-4o for the Data Generation Purpose.



	Reasoning	Knowledge	Conversation	Creativity	Personality	Storytelling	Empathy
LaMDA	0.84	0.69	1.0	0.53	0.85	0.58	0.94
ChatGPT	0.74	0.82	0.92	0.77	0.72	0.74	0.7
GPT-3	0.87	0.86	0.72	0.75	0.66	0.72	0.49
T5	0.7	0.6	0.19	0.51	0.1	0.36	0.04
PaLM	0.76	0.56	0.21	0.24	0.21	0.18	0.17
BLOOM	0.48	0.35	0.29	0.36	0.15	0.18	0.24
Turing-NLG	0.56	0.42	0.29	0.07	0.16	0.07	0.0

How popular LLMs score along human cognitive skills (source: semantic embedding analysis of ca. 400k AI-related online texts since 2021)

Figure 1: Comparision of LLMs. (ref. towarsdatascience.com)

3. **Fine Tuning:** Fine-tune the LLM response using advanced prompting techniques.

   For zero-shot prompting

   (a) The first level of prompting Used is Context Prompt. It defines the behaviour and role of the LLM

   (b) The second level of prompting used is Data Generation Prompt that generates our required dataset.

(c) The third level of prompting is Diversifying Prompt that brings in more diversification in the generated review content. With these level of prompting we can ensure a diverse and human like dataset for product reviews.

For few-shot prompting

(a) The first level of prompting is Context Prompt. It defines the behaviour and role of the LLM and also gives a sample of real world data to make the LLM familiar with it and make it capable of obtaining more real like dataset.

(b) The second level of prompting is Constraint Prompt that restricts the LLM from just copying and modifying provided example.

(c) The third level of prompting is Data Generation Prompt that generates the required data.

(d) The fourth level of prompting used is Diversifying Prompt that brings in more diversification in the generated reviews.

4. **Synthetic Data Generation:** Use the advanced prompting techniques to generate new reviews, ensuring a varied output.

5. **Data Augmentation:** Augment the Data Generated using various techniques to ensure a diverse and human generated like DataSet.

Questions and Answers

1. Why was the model/architecture used? The transformer architecture, particularly the GPT model, was chosen for its capabilities in natural language understanding and generation. Its self-attention mechanism allows it to capture context effectively, producing coherent and contextually relevant sentences. This results in synthetic reviews that are not only grammatically correct but also reflect the tone and style of actual user reviews. Also, as compared to other transformer based models, GPT-4o was found to be more reliable and cost effective model so it has been used for the generation purpose.

2. What were the different factors considered for generating this dataset? (Length, topic diversity, etc.) Several factors were considered while generating the synthetic dataset:

- **Length of Reviews:** An upper limit of 50 words was put for the content generation so that it as short and concise and imitates the real worls user reviews.
- **Topic Diversity:** Ensuring the dataset covers a range of topics (e.g., product categories, services, experiences) to reflect the diversity found in real reviews.
- **Sentiment Variation:** Incorporating different sentiments (positive, negative, neutral) to create a balanced dataset that mirrors the diversity of opinions expressed in real reviews.
- **Language and Style:** Maintaining a consistent language style that resembles the original dataset while allowing for variations in phrasing and vocabulary.

Below are some of the results from the code accompanying the above explanations:

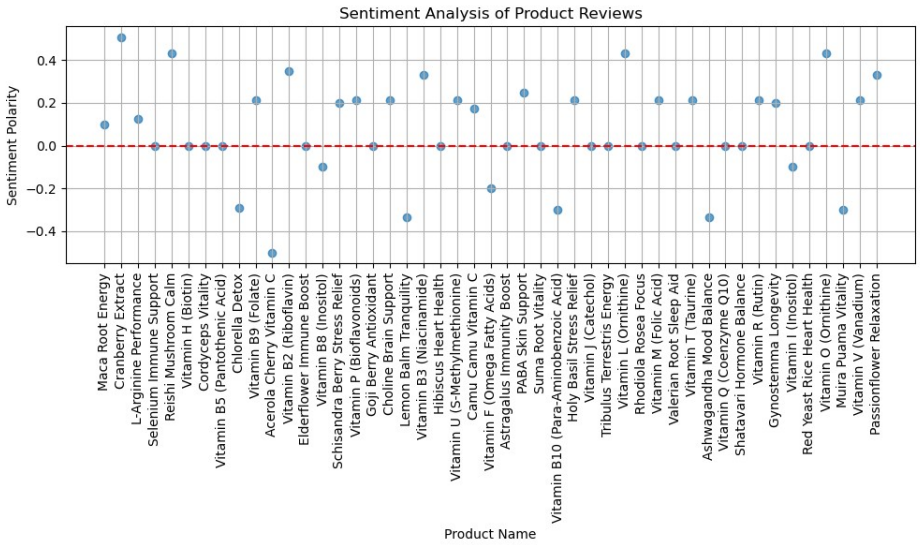


Figure 2: Sentiment Diversity with Zero Shot Prompting

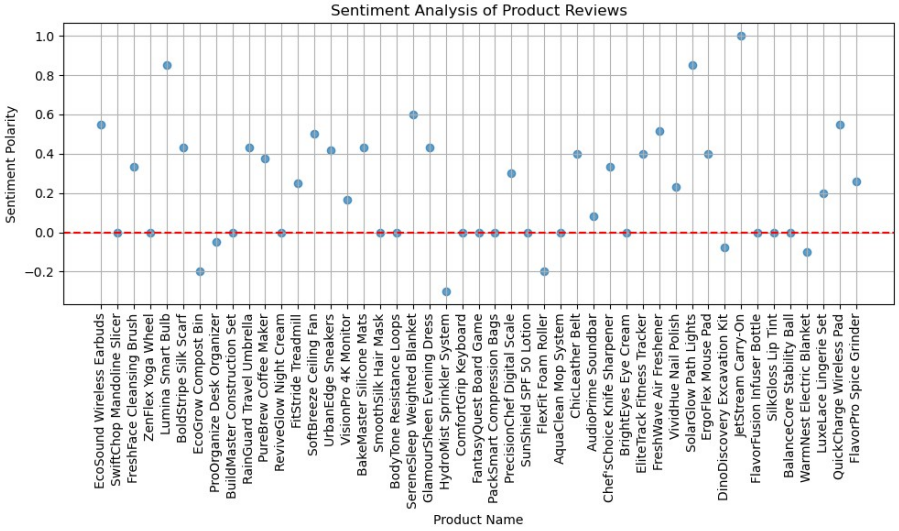


Figure 3: Sentiment Diversity with Few Shot Prompting

3. How do we measure the efficacy of a synthetic dataset? The efficacy of a synthetic dataset can be measured using the following metrics:

- **Diversity Metrics:** We measure the Remote clique Score and Chamfer Distance Score as the metrics to compute the diversity of the dataet as mentioned in this reference : arx and arxiv
  we can further performer deeper level of comparision using advanced metrics as mentioned in the above references but it has not been covered in the current version of analysis. The diversity scores for the generated and original dataset is shown below:

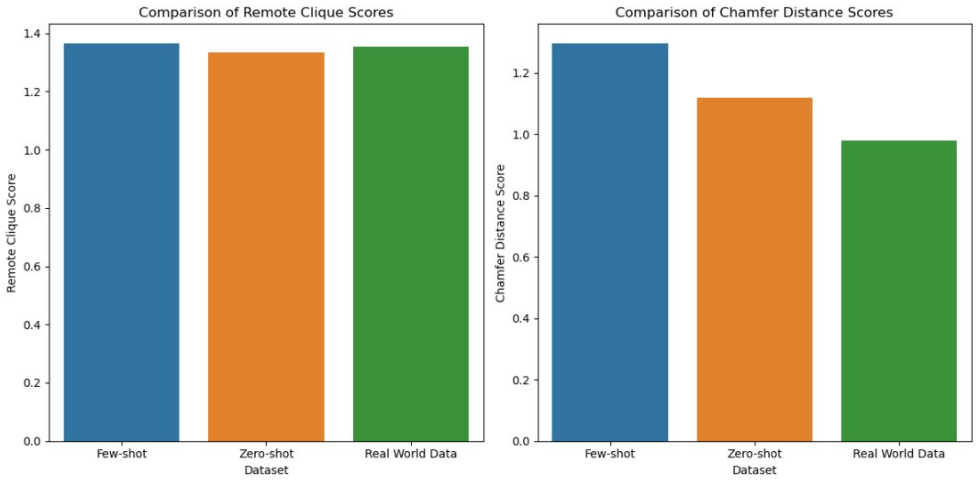


Figure 4: Diversity Scores for Zero-Shot, Few-Shot Generated Data and the real world review data

4. How do we ensure the synthetic dataset one generates is inspired from a source dataset but not an exact replica? To ensure originality while still being inspired by the source dataset, we employed the following strategies:

- **Data Augmentation:** Augmenting the original dataset and data generated through various methods to ensure diversity and uniqueness in the reviews.
- **Constraint Prompt:** Introducing the constraint prompt for data generation that restricts the llm to replicate the existing data provided to it.

5. What were the top challenges in solving for this problem statement? The main challenges encountered included:

- **Balancing Diversity and Coherence:** Achieving a balance between producing diverse reviews while maintaining relevance and uniqueness.
- **Quality Control:** Ensuring the generated reviews are realistic and high-quality, which often required iterative fine-tuning of the prompts and using advanced prompting techniques.
- **Evaluation Complexity:** Developing effective metrics and methods for evaluating the synthetic dataset's quality, which is inherently subjective.

Conclusion The generation of a synthetic dataset of reviews is a achievable task, requiring careful consideration of various factors like uniqueness, diversity, human like text, text length etc. Through the use of advanced LLMs and prompting methodologies, we can create datasets that are diverse, high-quality, and useful for various applications in natural language processing.