ITCS-6190
Cloud Computing for Data Analytics
LLM - Detect AI Generated Text

Team Members

Apoorve Bhargava
Soumik Paul
Juhi Jadhav
Raj Kumar
Sai Kumar Reddy, Bommareddy

Dataset: LLM - Detect AI Generated Text

AWS Services Used
AWS S3
AWS Athena
AWS Glue
AWS Quicksight
AWS Sagemaker

About the data

The LLM (Large Language Model) Detect AI Generated Text dataset is likely designed for training models to identify and differentiate between human-generated and AI-generated text. It encompasses a diverse collection of text samples, featuring instances of both human and machine-generated content. This dataset aims to enhance the capabilities of language models in discerning the unique patterns and characteristics associated with AI-generated text, aiding in the development of effective detection algorithms. Researchers and developers use such datasets to advance the field of natural language processing by creating models that can reliably identify content produced by sophisticated language models like GPT (Generative Pre-trained Transformer).

There are 4 CSV files in the dataset with respective columns

1. **Train_Essays.csv: Id, prompt_id, text, generated**
2. **Train_Prompts.csv: prompt_id, prompt_name, instructions, source_text**
3. **Test_essays.csv : id, prompt_id, text**
4. **Sample_submission.csv : id, generated**

Creation of Bucket and storing the dataset into the bucket

Data Preparation using AWS Glue

Step 1: AWS Glue is used to fetch the data from S3 to AWS Glue

✓ Crawler successfully starting
The following crawler is now starting: "crawlergroup7"   ✕

AWS Glue > Crawlers > crawlergroup7

crawlergroup7

Last updated (UTC)
November 19, 2023 at 20:37:04   ↻   Run crawler   Edit   Delete

Crawler properties

Name	IAM role	Database	State
crawlergroup7	LabRole ↗	group7db	READY
Description	Security configuration	Lake Formation configuration	Table prefix
-	-	-	-
Maximum table threshold			
-			

▶ Advanced settings

Crawler runs | Schedule | Data sources | Classifiers | Tags

Crawler runs (1)
The list of crawler runs for this crawler.   ↻   Stop run   View CloudWatch logs ↗   View run details

Q Filter data		🗓 Filter by a date and time range				‹ 1 › ⚙
Start time (UTC) ▲	End time (UTC) ▽	Current/last duration ▽	Status ▽	DPU hours ▽	Table changes	▽
○ November 19, 2023 at 20:37:	November 19, 2023 at 20:38:	55 s	✓ Completed	-	4 table changes, 0 partition changes	

AWS Glue > Crawlers

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (1/1) Info
View and manage all available crawlers.

Last updated (UTC)
November 19, 2023 at 20:39:20   ↻   Action ▼   Run   Create crawler

Q Filter crawlers					‹ 1 › ⚙
☑ Name ▽	State ▽	Schedule	Last run ▽	Last run timestamp ▽	Log	Table changes from la...
☑ crawlergroup7	✓ Ready		✓ Succeeded	November 19, 2023 at 2...	View log ↗	4 created

Data Transformation using Visual ETL

Step 2: Preview the data on AWS Athena
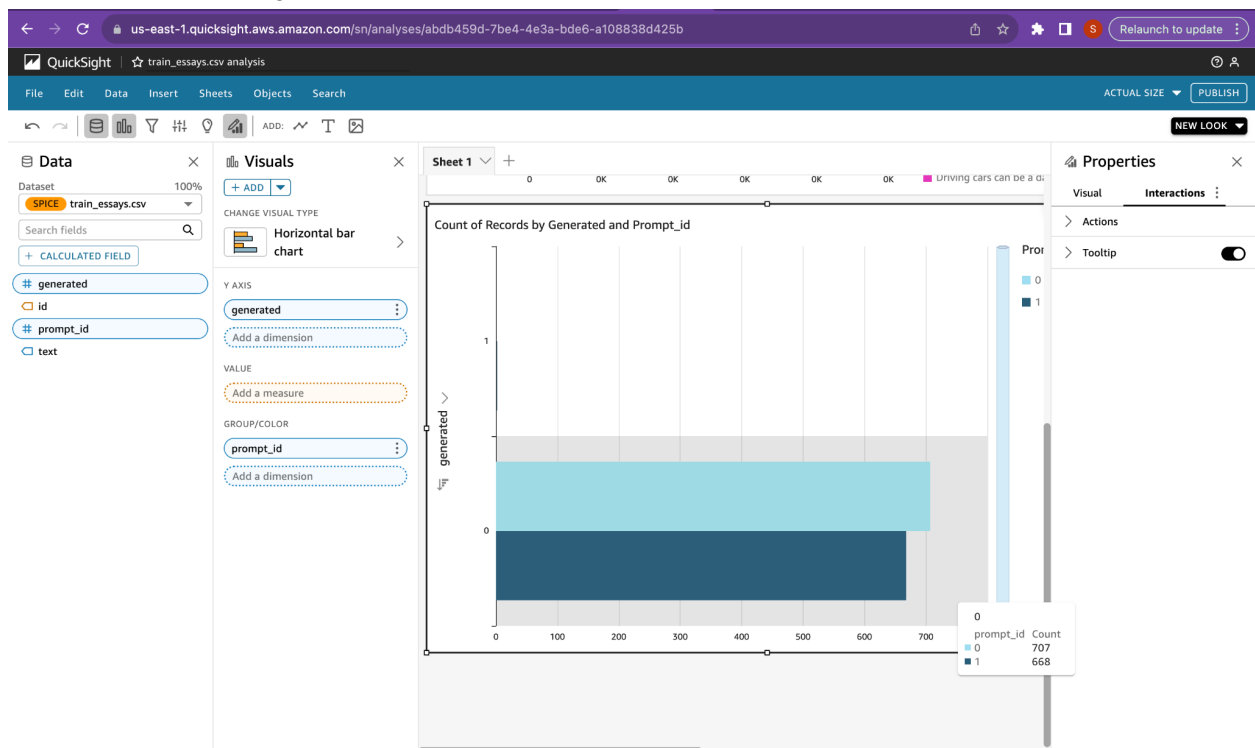Run a few SQL queries on AWS Athena to understand the datasets.

Data Visualization on Quick Sight
The dataset has been visualized on AWS Quick Sight to get a better understanding of the data

Count of Records by IDs and Text
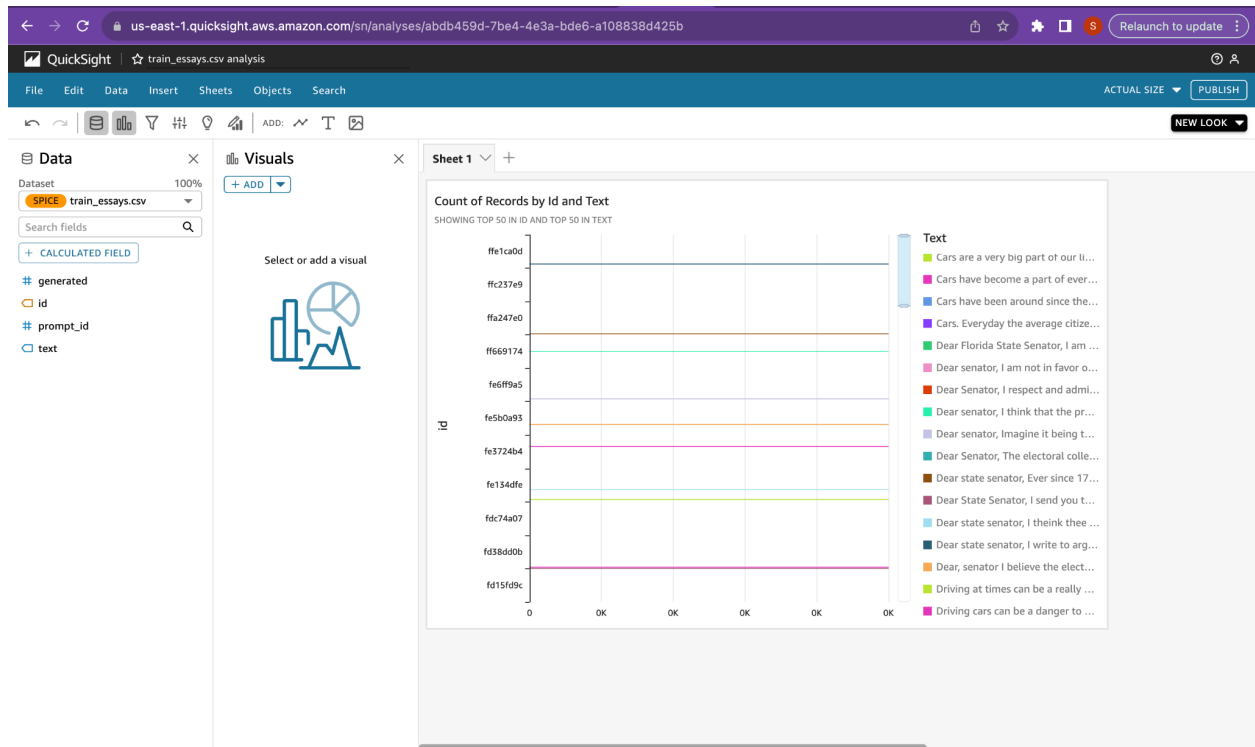
Count of Records by Generated and Prompt_Ids

Count of Records by Generated Text, and Prompt_Ids