

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

Amazon S3

Buckets

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

Storage Lens groups New

AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

Upload succeeded

View details below.

Upload: status

Close

The information below will no longer be available after you navigate away from this page.

Summary

Destination

s3://mybucketgroup7

Succeeded

4 files, 4.2 MB (100.00%)

Failed

0 files, 0 B (0%)

Files and folders

Configuration

Files and folders (4 Total, 4.2 MB)

Find by name

Name	Folder	Type	Size	Status	Error
sample_submission.csv	-	text/csv	52.0 B	Succeeded	-
test_essays.csv	-	text/csv	90.0 B	Succeeded	-
train_prompts.csv	-	text/csv	26.1 KB	Succeeded	-
train_essays.csv	-	text/csv	4.2 MB	Succeeded	-

CloudShell

Feedback

© 2023, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

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

aws

Services

Search

[Option+S]

🔍

⬆️

⬆️

⬆️

N. Virginia

vodabs/user2103908=spaul13@uncc.edu @ 5813-8714-1831

AWS Glue

Getting started

ETL Jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

ETL Jobs

Visual ETL

Notebooks

Job run monitoring

Interactive Sessions

Data classification tools

Sensitive data detection

Record Matching

Triggers

Workflows (orchestration)

Blueprints

Security configurations

Legacy pages

What's New

AWS Glue > Databases

Databases (1/1)

A database is a set of associated table definitions, organized into a logical group.

Filter databases

<input checked="" type="checkbox"/>	Name	Description	Location URI	Created on (UTC)
<input checked="" type="checkbox"/>	group7db	-	-	November 19, 2023 at 20:36:35

Last updated (UTC)
November 19, 2023 at 20:36:37

Edit

Delete

Add database

aws

Services

Search

[Option+S]

🔍

⬆️

⬆️

⬆️

N. Virginia

vodabs/user2103908=spaul13@uncc.edu @ 5813-8714-1831

AWS Glue

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

ETL Jobs

Visual ETL

Notebooks

Job run monitoring

Interactive Sessions

Data classification tools

Sensitive data detection

Record Matching

Triggers

Workflows (orchestration)

Blueprints

Security configurations

Legacy pages

What's New

You can now create Apache Iceberg tables in the AWS Glue Data Catalog. To learn more, visit the [documentation](#).

Create table

AWS Glue > Databases > group7db

group7db

Last updated (UTC)
November 19, 2023 at 20:39:46

Edit

Delete

Database properties

Name	group7db	Description	-	Location	-	Created on (UTC)	November 19, 2023 at 20:36:35
------	----------	-------------	---	----------	---	------------------	-------------------------------

Tables (4)

View and manage all available tables.

Filter tables

<input type="checkbox"/>	Name	Database	Location	Classification	Deprecated	View data	Data quality
<input type="checkbox"/>	sample_submission_csv	group7db	s3://mybucketgroup7/sample	CSV	-	Table data	View data quality
<input type="checkbox"/>	test_essays_csv	group7db	s3://mybucketgroup7/test_ess	CSV	-	Table data	View data quality
<input type="checkbox"/>	train_essays_csv	group7db	s3://mybucketgroup7/train_ess	CSV	-	Table data	View data quality
<input type="checkbox"/>	train_prompts_csv	group7db	s3://mybucketgroup7/train_prom	CSV	-	Table data	View data quality

Last updated (UTC)
November 19, 2023 at 20:39:46

Delete

Add tables using crawler

Add table

Services Search [Option+5]

- [Notebooks](#)
- [Job run monitoring](#)
- [Data Catalog tables](#)
- [Data connections](#)
- [Workflows \(orchestration\)](#)
- Data Catalog**
 - Databases
 - Tables
 - Stream schema registries
 - Schemas
 - Connections
 - Crawlers**
 - Classifiers
 - Catalog settings
- Data Integration and ETL**
 - ETL jobs
 - Visual ETL
 - Notebooks
 - Job run monitoring
 - Interactive Sessions
 - Data classification tools
 - Sensitive data detection
 - Record Matching
 - Triggers
 - Workflows (orchestration)
 - Blueprints
 - Security configurations
- Legacy pages**
 - What's New
 - Documentation
 - AWS Marketplace

☐ Enable compact mode
☒ Enable new navigation

[CloudShell](#) [Feedback](#)

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
 Refresh
Run crawler
Edit
Delete

Crawler properties

Name crawlergroup7	IAM role LabRole	Database group7db	State 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

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
<input type="radio"/>	November 19, 2023 at 20:37:	November 19, 2023 at 20:38:	55 s	Completed	-	4 table changes, 0 partition changes

© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

[illegible]

Step 1: Create a table in the AWS Glue Data Catalog

You can now create Apache Iceberg tables in the AWS Glue Data Catalog. To learn more, visit the [documentation](#).

Tables

A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Tables (4)

View and manage all available tables.

Filter tables

Name	Database	Location	Classification	Deprecated	View data	Data quality
sample_submission_csv	group7db	s3://mybucketgroup7/sample_submission_csv	CSV	-	Table data	View data quality
test_essays_csv	group7db	s3://mybucketgroup7/test_essays_csv	CSV	-	Table data	View data quality
train_essays_csv	group7db	s3://mybucketgroup7/train_essays_csv	CSV	-	Table data	View data quality
train_prompts_csv	group7db	s3://mybucketgroup7/train_prompts_csv	CSV	-	Table data	View data quality

© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Step 2: Preview the data on AWS Athena

Run a few SQL queries on AWS Athena to understand the datasets.

Amazon Athena > Query editor

Editor Recent queries Saved queries Settings

Workgroup: primary

Athena now supports typeahead code suggestions to speed up SQL query development. Typeahead suggestions are turned on by default. You can change this setting in query editor preferences. [Edit preferences](#)

Data

Data source: AwsDataCatalog

Database: newdb07

Tables and views

Filter tables and views

Tables (1)

mybucketgroup7

id: string

prompt_id: bigint

text: string

generated: bigint

Views (0)

SQL Ln 1, Col 51

```
SELECT * FROM "newdb07"."mybucketgroup7" limit 10;
```

[Run again](#) [Explain](#) [Cancel](#) [Clear](#) [Create](#)

Query results **Query stats**

Completed Time in queue: 112 ms Run time: 498 ms Data scanned: 0.53 KB

Results (10)

Search rows

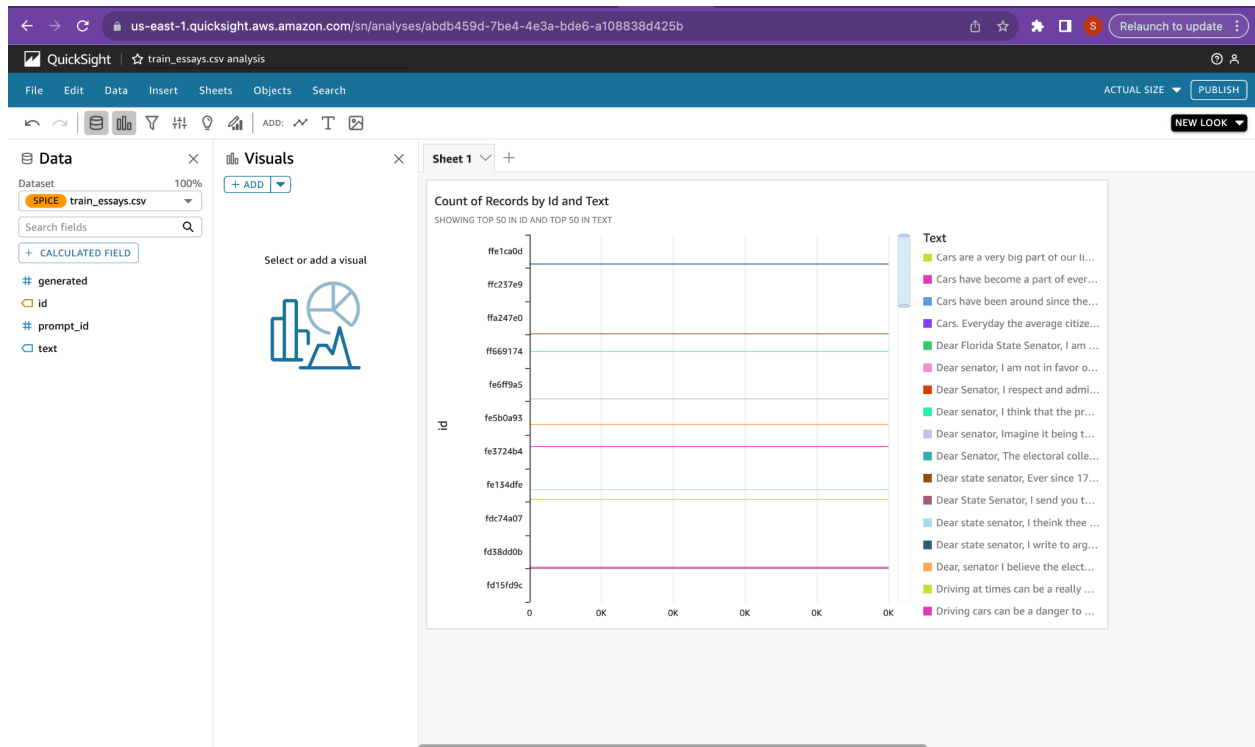
#	id	prompt_id	text	generated
1	"0059830"		"Cars. Cars have been around since they became famous in the 1900s"	
2	"			
3	"In like matter of this"		" ""in German Suburb"	
4	"			
5	"Likewise"		" ""Carfree day is spinning into a big hit in Bogota"	
6	"			

© 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

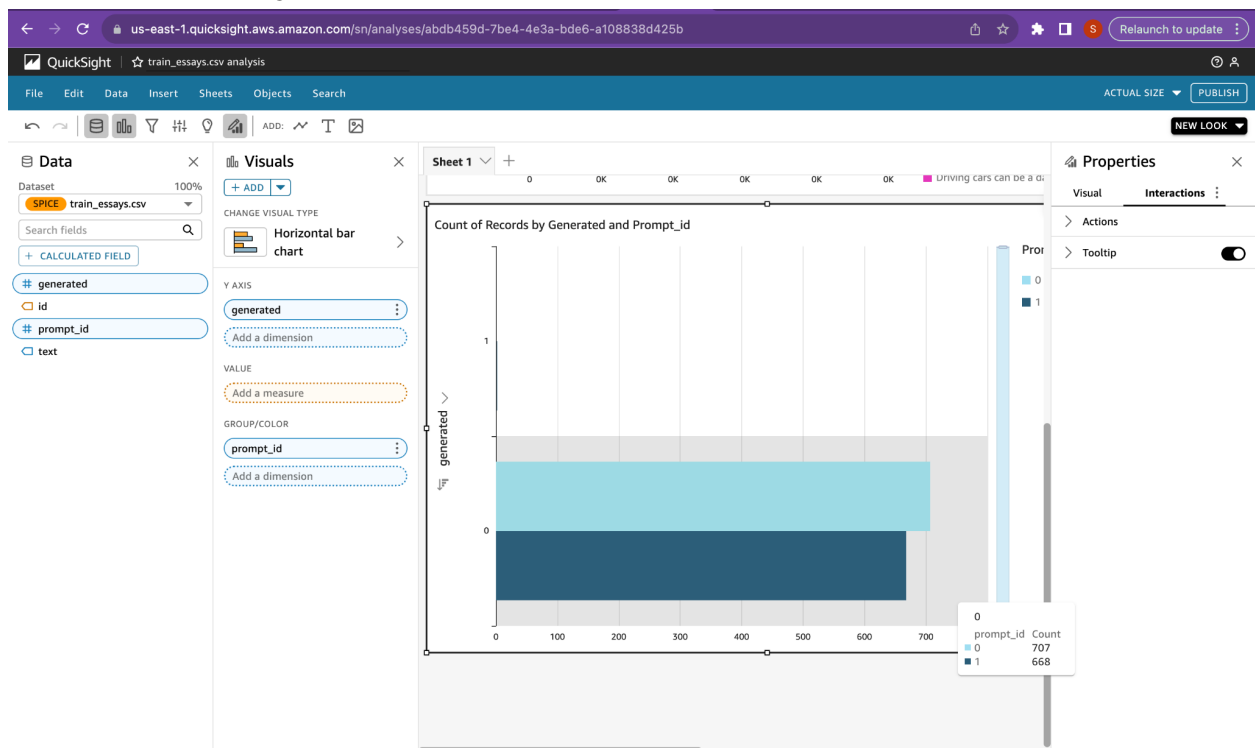
Data Visualization on Quick Sight

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

