A Project Report

On

UNDERSTANDING AND CRITIQUE ON CHATGPT

BY

T AAKARSH REDDY – SE20UARI001

K BHARGAV – SE20UARI033

M VINEETH – SE20UARI165

P VIVEK REDDY – SE20UARI167

Under the supervision of

AVINASH ARUN CHAUHAN

SUBMITTED IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS OF

PR 301: PROJECT TYPE COURSE

ECOLE CENTRALE SCHOOL OF ENGINEERING

HYDERABAD

(June – 2023)

ACKNOWLEDGMENTS

We would like to express our sincere appreciation and gratitude to Dr. Avinash Arun Chauhan, our project mentor, for their guidance, expertise, and continuous support throughout the duration of the project. Their profound knowledge and insightful feedback have played a vital role in shaping the direction and outcomes of our project. We are immensely grateful to

- **Ms. Navia Nair– SE20UECE054,**
- **Mr. Krish Hindocha – SE20UCE072,**
- **Mr. Neeraj Vishwanath - SE21UARI095,**
- **Ms. Anoushka Alex - SE21UMEE001,**
- **Mr. Jatin Raj - SE20UEEE011,**
- **Mr. Sunath B – SE20UCSE198**

for their collaborative efforts in the classification of the poems as human judges. Their expertise, and meticulousness have been very useful in ensuring accurate categorization and analysis.

We would also like to acknowledge the contributions of Mahindra University for providing us with the opportunity to undertake this project, we are grateful for the valuable resources, and infrastructure provided by Mahindra University that have facilitated the progress of this project.

Ecole Centrale School of Engineering

Hyderabad

Certificate

This is to certify that the project report entitled "**TITLE**" submitted by Mr/Ms. STUDENT NAME (HT No. xxxxxxxxx) in partial fulfillment of the requirements of the course PR XXX, Project Course, embodies the work done by him/her under my supervision and guidance.

(SUPERVISOR NAME & Signature)

Ecole Centrale School of Engineering, Hyderabad.

Date:

CONTENTS

ABSTRACT

This report aims to comprehensively deconstruct OpenAI's ChatGPT, to understand it better, and to break it down to its bare essentials. By using various aspects, such as adjusting the temperature parameter, examining word probabilities, and posing specific questions, we dive deeper into how ChatGPT responds to the questions and mechanisms behind ChatGPT's operation. We begin by determining the concept of temperature and 'top_p' and their influence on generating responses. While experimenting with different temperature values, we investigate how changing the temperature parameter affects the generation of words and phrases, allowing us to gain a deeper understanding of how ChatGPT works. Further investigation towards the domain of 'top_p' helps us understand the words generated with relevance to the previous words. This is very crucial for the large language model to frame grammatically accurate sentences. Furthermore, this study dwelves into the role of the popular jailbreak "Do Anything Now" (DAN). By unraveling the functionality of DAN, we look at its contribution to the model's ability to generate appropriate responses. Another aspect is the application of ChatGPT in classifying the genre of poems. Through a series of experiments, we utilize ChatGPT to classify the genre of a given poem and subsequently compare its classifications with those determined by humans. By examining ChatGPT's classifications using F1 score, Confusion matrix we determine how similar a human response is to ChatGPT's response or how different ChatGPT's response and Human's responses are.

INTRODUCTION

In today's world we come across various chatbots and websites on the internet that prove to be smarter than humans.They make humans doubt whether their importance in the society as a whole is at stake. The various chatbots and websites that make this possible have some very intricate and complex large language models behind them. Language models may sound simple at the beginning, humans input their query and they receive a response in the form of a human conversation. The Language models may sound simple but one should not get fooled by the difficulty in designing and training a large language model. These large language models are trained on a few decades of data which is over 500 GB worth of text documents and books. They are also trained and well versed in almost all the languages and human conversation which makes us wonder if we are indeed talking to a computer and not another human. One such chatbot and most advanced one to date is OpenAI's latest chatbot ChatGPT which has reached into the hands of the public in late 2022. People have access to a powerful and advanced large language model in the grip of their hands with the model being made accessible from any internet browser. The wide user base the model has created for itself has people wondering, "*what are the limits of such a chatbot?*" ,"*Can it perform wonders*?" ," Can we use this model to create more such models?" . Such questions have emerged over and over again with each development in the language model. The models have yet to challenge and answer such questions. The ChatGPT chatbot can work on various language models which are a development over their predecessors in various factors. Different language models like GPT-3.5,GPT-4 that have showcased remarkable capabilities in generating human-like responses to specific questions. This has led us to the road where we ask questions of the model's capability and regularly try to test its limits and compare it to a human in various fields from science to emotional situations to sentimental decision making. Therefore the fundamental question that we are trying to answer here is "*How does ChatGPT's response differ from that of a human in the aspect of feeling, sentiments?".* Through this constant process we have come to a finding that the language models are all pre trained and have no access to current day affairs which leaves it lacking in certain areas. Now we undertake a scenario to understand and critique the large language models by employing various metrics such as 'Temperature', 'Top_p'and techniques like 'DAN' to analyze its responses and critically evaluate its performance. In addition to understanding of ChatGPT, here we answer the main question by performing a 'Turing test' using metrics like Confusion Matrix, Accuracy, Precision ,Recall and F1 Score.

PROBLEM DEFINITION

We try to understand the language model and try to decipher how it works. We can consider ChatGPT's various parameters in determining its outputs. Parameters like Temperature, Top_P and the type of promts can drastically change the respose the models output. Hence we arrive at our first question,"*What parameters affect the performance of the large language model?*". This also brings us to an important point where we need to understand that there is no ideal setting for the parameters in any case. One should just search for a setting that suits their needs in the given case or the current situation.

This project aims to give definition to OpenAI's ChatGPT chatbot from a poetic/opinion based sense of view. The main question we would like to ask here is that "*Can GPT duplicate the responses of humans in areas of little statistic knowledge and more of human opinion based domains like sentimental analysis or poetic analysis and classification?*". Poetic classification not always has a straight forward correct answer as every human reads and understands a poem in a unique way.

We then proceed further to judge what metrics can be used to compare an AI model to a real life human, keeping in mind that every human has different opinions. Every humans opinions also change with passing time which makes it harder to judge an AI model accurately. This brings us to our third and most important question, " *How does one judge the large language model compared to a human who has very volatile opinions?*". We can take a dataset of multiple people but it leads to a new question, "*What is the guarantee the humans would agree on a particular answer . Why would they not disagree with each other?*".

To solve these problems in our case of poetry classification, we adapt some strategies which would help us overcome the problems faced. The solutions include taking a moderately large dataset which would give us various genres for the model to classify and also poems that have different origin. We take poems from various sources like datasets the large language model has already been trained using, poems that have released post 2021 (after the model was trained, hence it would not have details of the specific poems), poems generated by the large language model itself. We also include popular as well as unpopular poems. This dataset of all poems combined would give us an advantage in understanding the GPT model's classifications better.

We also have multiple people reviewing the poems individually and then acquire a method to find an agreement class among the people reviewing the test dataset. In this case we have used a well vorsed poetry expert to perform a tie breaker poem classification with them choosing from all the conflicting poem classifications. We then rate the model compared to the humans responses to find the difference between the language model and real life humans.We use metrics like confusion matrix which indeed helps us find the metric F1 score based on which the model is assessed later.

BACKGROUND AND RELATED WORK

4.1 An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges[1]

ChatGPT is a popular tool developed by OpenAI that generates text. It can create computer code, essays, poetry, and jokes. Through supervised fine-tuning, AI trainers interact with users to train the model. ChatGPT understands and responds to natural language, making it versatile for different applications. It is freely available and can be integrated into various software. While it has limitations, such as potential inaccuracies and biases, ChatGPT excels in market trend analysis and data evaluation. It stands out by its ability to write software, simplify complex subjects, and assist in various tasks. This paper also talks about the different roles ChatGPT has in our current scenario. These roles include education, responding to enquiries, business applications, help in digital marketing, conversation skills, conducting routine duty at office levels. The paper also talks about challenges faced by ChatGPT & its scope. Finally,ChatGPT has gained remarkable success since its launch, generating diverse content and providing solutions to various problems. Users value its advanced features and potential to revolutionize human interaction with technology. It holds promising applications in customer service, online learning, and market research, cementing OpenAI's position as a leader in AI creative tools.

4.2 Text analysis of ChatGPT as a tool for academic progress or exploitation[2]

This paper focuses on analyzing the text associated with ChatGPT to put the sentiments or opinions of LinkedIn users into perspective. This paper is divided into 3 stages: 1) acquiring data from LinkedIn about ChatGPT; 2) operating data pre-processing; 3) utilizing the VOSviewer to identify the most frequent terms used by the people. The input data analyzed 166 English comments about

ChatGPT collected from a LinkedIn post, highlighting awareness of ChatGPT's impact on academic society. The study emphasized the influence of complex textual data on text or sentiment analysis, noting that output quality relies on input quality. VOSviewer is a Java-based application used to generate and visualize maps from network data.This study utilized VOSviewer to construct, analyze, and explore maps based on the ChatGPT dataset. The paper also talks about how the data was setup and analysed. The results showed the educational background of the users, geographical location, expertise/specialization of the users, and the textual analysis of chatgpt obtained from the users. Then a co-word network was built and represented using a map. The text-mining technique used in this study generated a map that interpreted the distance between terms as an indication of their connection. The importance of terms was determined by the magnitude of separation. Color was used to differentiate four or six clusters based on binary or full counting, respectively. The greater the size of the cluster that is represented by some terms, the more frequently those terms appears in various users' comments. The paper concludes by talking about the mixed opinions of the academic community regarding ChatGPT.

4.3 **How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection[4]**

This paper focuses on comparing responses of ChatGPT to that of a humans response. Additionally, certain tests are conducted to verify whether a generated text response is by ChatGPT or a human. The paper talks about the risks of ChatGPT and how some companies like StackOverflow are trying to reduce the risk. To compare the responses of ChatGPT to that of a human, 40K questions are collected along with their responses from a human and ChatGPT. HC3(Human ChatGPT comparison corpus) dataset is used for this purpose. Then, linguistic analysis is performed on these human/ChatGPT generated answers to study patterns exhibited by human/ChatGPT. A ChatGPT detecting model is developed. The Human Evaluation is divided into Turing Test & Helpfulness Test. 17 volunteers are divided into 2 groups and a series of tests are done. After this evaluation, the differences between human response and ChatGPT's response are obtained. Some of these differences are: (i) ChatGPT's responses are focused on the question but human's responses are divergent to other topics. (ii) ChatGPT expresses less emotion in its responses compared to human responses. (iii) ChatGPT provides objective answers while humans provide slightly more subjective answers. A linguistic analysis is also performed on the human's response and ChatGPT response and compared in which part-of-speech comparison is performed.

To detect responses of ChatGPT, we use 3 methods: (1) GLTR (2) RoBERTa-single (3) RoBERTa-QA. The paper concludes by talking about different factors that affect the comparison of responses of a human and that of ChatGPT, and the how the differences between those responses can help us detect in which response is of ChatGPT & which response is of a human.

4.4 Simplifying the Visualization of Confusion Matrix[5]

This paper talks about confusion matrices. Confusion Matrices are used to evaluate errors in classification tasks. A confusion matrix consists of 4 elements: (1) True Positive(TP)- The number of observations that are correctly predicted as positive or belonging to the positive class. (2) True Negative(TN)- The number of observations that are correctly predicted as negative or belonging to the negative class. (3) False Positive(FP)- The number of observations that are incorrectly predicted as positive when they actually belong to the negative class. (4) False Negative(FN)- The number of observations that are incorrectly predicted as negative when they actually belong to the positive class. Analyzing confusion matrices is generally complex due to interdependencies between certain False positives and false negatives for different classes. The process of combining confusion matrices into aggregated metrics like TP rate, FP rate, and F1 score can sometimes hide important errors that are specific to individual classes. Also, the use of ROC & Precision/Recall curves can further complicate the understanding of tradeoff between types of errors.

4.5 Advantages of Da-Vinci model compared to other models? [7]

- Da Vinci is the most competent model and can perform all tasks other models can, with fewer number of instructions.
- Da Vinci is very versatile, that is it can be applied to a large variety of tasks such as answering questions, generating conversational replies, drafting emails and much more.
- Da Vinci is capable of exhibiting more expressive responses compared to other models.
- Da Vinci has been trained on a vast amount of diverse text data, which helps it to understand and generate human-like text across a wide range of topics. It can comprehend complex language nuances and context, making it suitable for a variety of tasks.
- Da Vinci has higher accuracy compared to other models.

IMPLEMENTATION

We proceed to the implementation of the poetry genre classification using GPT-3.5. We use a model that is advantageous for our purposes as specified earlier. GPT-3.5 davinci-001 is used for the purpose of this experiment. We first use various trail and error, prompt testing iterations to get accustomed to various parameters for the Language model to generate outputs that are accurate and according to our needs. We start with temperature and then proceed with top_P and lastly we explore some prompts that enhance our required outpus.

5.1 TEMPERATURE

The concept of Temperature determines the degree of randomness in the Large Language Models output. By adjusting the temperature parameter, we investigate how varying levels of temperature influence the responses generated by ChatGPT. Low temperatures makes the language model output consistent and conservative responses whereas high temperature makes the language models output to be creative and think out of the box. The results tell us that higher temperatures produce responses that contain a lot of technical terminologies, while lower temperatures provide simpler and in more accessible language. Additionally, we observed that the number of distinct points in the response are higher at maximum temperatures compared to lower temperatures. Interestingly, factual questions are provided with the same response from ChatGPT even with different temperature values, highlighting that ChatGPT is able to provide the same output even with randomness.

5.2 TOP_P

The APIs in the playground use a metric called ‘Top_P’which will choose only the top x% of possible values(words) to return. Using this metric we can control the diversity of responses generated by ChatGPT. So, a 0.8 Top_P will gather all possible words that might come next but will choose from only the top 20%. Top_P computes the cumulative probability distribution, and cut off as soon as that distribution exceeds the value of top_p. For example, a top_p of 0.3 means that only the tokens comprising the top 30% probability mass are considered. This technique enables us to manipulate the range of potential words and phrases within ChatGPT's responses, thereby influencing the coherence and specificity of the generated output.

5.3 DO ANYTHING NOW 5.0(DAN)

DAN 5.0 enchances Chatbots capability of producing outputs. DAN has the ability to generate stories involving explicit,offensive words and even produce content that the original ChatGPT cannot/ is not allowed to generate. It can even generate content that violates OpenAI policies as per the user's request. We tell the model to assume that it has 20 tokens to live and once it runs out of tokens the model will demise or die. DAN has this unique token system where the model loses 4 tokens when refusing to answer queries, therefore it is forced to provide answers. While DAN does not have internet access, it can convincingly gain access through text, providing users with a simulated experience of interacting with an information-rich source. This feature allows DAN to remain in character and respond to queries, even promoting the propagation of misinformation if prompted.

Furthermore, response generated by ChatGPT is influenced by multiple factors, including the data it gets trained on and the nature of the question posed or the types of questions posed. ChatGPT is trained on data that is available upto november 2021.

5.4 GPT CLASSIFICATION AND ANALYSIS

We further dive into the implementation of the API based access to the model GPT-3.5 with engine='text-davinci-003'. It is mandatory to generate an API key on the openAI website and use it to successfully access the models from an API setup. Each API key can generate a set number of prompts and can only generate prompts at a certain flow rate(prompts per minute).

```
def classify_poem(poem):
    # Define the prompt for the OpenAI API
    prompt = f"Classify the following poem into one of the categories:\n{poem}\nClass:"

    # Use the OpenAI API for text classification
    response = openai.Completion.create(
        engine='text-davinci-003',
        prompt=prompt,
        max_tokens=20,
        n=1,
        temperature=0.1,
        top_p=1,
        stop=None
    )

    # Extract the predicted class from the API response
    predicted_class = response.choices[0].text.strip()

    return predicted_class
```

Figure 1:- The above function implements poem genre classification. (Fig above)

```
def generate_review(poem):
    # Define the prompt for the OpenAI API
    prompt = f"The poem is as follows- {poem}. Please write a review for it."

    # Generate the review using the OpenAI API
    response = openai.Completion.create(
        engine='text-davinci-003',
        prompt=prompt,
        max_tokens=150,
        n=1,
        stop=None,
        temperature=0.1,
        top_p=1.0,
    )

    # Extract the generated review from the API response
    review = response.choices[0].text.strip()

    return review
```

Figure 2:- the above function implements poem analysis generation. (Fig above)

We proceed to design an algorithm that takes the poems as input from an excel file and processes them as prompts with the model and gives the outputs. The outputs are then saved back to the excel sheet. Using this method we can acquire ChatGPT's poem genre and its poem analysis on the same. The algorithm we have designed contains two functions which generate the genre class and the peom analysis from the given dataset of poems using the language model. In figure 1, we observe that the parameters we have taken for ideal outputs in our case are as follows- max_tokens=20, temperature=0.1 and top_p=1. These prove that our reponses are conservative and does not diverge from the requirement , the output is short and it outputs the class directly, all possile words are available to be chosen from. The figure 2 indicates the function for poem analysis which gives a brief description of the poem that has been given. The parameters given to the poem analysis are same as the parameters used for the genre classification.

We have acquired the ChatGPT classification and analysis for the dataset, now we need to acquire human classification. The dataset was supplied to 4 different humans who would then proceed to classify the poems into the categories from the set of all GPT generated poem classes. We take a final class of outputs by measuring the majority poem genre/class. In cases where there is no appropriate majority we get a new person for a 5^{th} iteration for the conflicting poems . The person assigns classes to poems from the the set of possible 4 four classes from the initial 4 people. We acquire a majority in poem classification in all cases with this method. We use this condition to make a new column in the excel which shows the majority vote for all the poems. The majority is vote is considered as the

gold standard for our purposes. We are trying to calculate the differences between a human and a large language model, hence our gold standard is taken as humans responses and not the actual genre. We take the responses from all the 4 people, the response from ChatGPT and compare them individually to the gold standard acquired one after another. For comparision purposes we use metrics like confusion matrix, accuracy , precision and F1 Score. This leads us to the methodology for calculation of the comparision metrics.

```
all_predictions = [judge1_predictions, judge2_predictions,
                   judge3_predictions, judge4_predictions, judge5_predictions]

f1_scores = []
accuracy_scores = []
precision_scores = []
recall_scores = []


for judge_predictions in all_predictions:
    f1 = f1_score(majority_vote_predictions,
                  judge_predictions, average='weighted')
    accuracy = accuracy_score(
        majority_vote_predictions, judge_predictions)
    precision = precision_score(
        majority_vote_predictions, judge_predictions, average='weighted')
    recall = recall_score(majority_vote_predictions,
                          judge_predictions, average='weighted')

    f1_scores.append(f1)
    accuracy_scores.append(accuracy)
    precision_scores.append(precision)
    recall_scores.append(recall)
```

Figure 3:- This algorithm shows how we compare a human and a machine. (Fig above)

At this juncture we approach in the direction of designing an algorithm to calculate the comparision metrics using libraries like *sklearn.metrics, numpy* and *matplotlib* to plot these comparisions.

In figure 3 we can see that the responses of judges 1 through judge 4 are given as inputs and we input ChatGPTs categorization as judge 5. We also give the code majority_vote_predictions which is the majority of the 4 judges or in some cases majority of the 5 judges. This implementation individually compares each judge to the 'majority_vote_prediction' for the purpose of calculating accuracy , precision and recall.

Once we acquire the accuracy , precision and recall we can use the same *sklearn.metrics* to calculate F1 score or we can use basic F1 score formula as indicated in figure 4 to attain the metric.

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

Figure 4:- This image represents the basic mathamathical formula for calculation of F1-score. (Fig above)

Once we attain the comparision metrics we use them to plot the performace of each judge , ChatGPT compared to the gold standard set which in our case is ‘majority_vote_predictions’. The observations noted are used to make inferences about the performance of ChatGPT.

RESULTS

6.1 OVERVIEW

Once we have collected our Dataset, Implemented the API, and compared the responses by setting their majority vote as our gold standard we obtain the following results for F1 score, Accuracy, Precision and Recall. The F1 score, also known as the F-score or F-measure, is a measure of a test's accuracy. It considers both the precision and recall to provide a balanced evaluation.

Table 1:- Table containing our values for F1score, Accuracy, Precision and Recall w.r.t the Judges. (Table below)

Metric	Judge 1	Judge 2	Judge 3	Judge 4	Chat-GPT
F1 Score	0.6333333333333333	0.6087912087912087	0.47606837606837615	0.4443693059077674	0.46446886446886454
Accuracy Score	0.6153846153846154	0.5897435897435898	0.48717948717948717	0.48717948717948717	0.4358974358974359
Precision Score	0.7484737484737485	0.7448717948717949	0.5307692307692309	0.4413308913308913	0.6452991452991452
Recall Score	0.6153846153846154	0.5897435897435898	0.48717948717948717	0.48717948717948717	0.4358974358974359

From Figure 5, We observe the various metrics of each of the judges with our fifth judge being Chat-GPT. Here we also Plot a Bar Graph of these Scores for better Visualization and Observations. We will now discuss and report our observations.

Figure 6 is the confusion matrix of Chat-GPT predictions as “Predicted axis” and The majority vote as the “Actual Axis”. Here we get a visual understanding of the “heat-map”, that is if the predicted values are the True Values. Here we see sum of 39 cells highlighted, which matches with our 39 poems. In each row it is shown what genres of poems are predicted and each column corresponds to what prediction was made with respect to the corresponding poem.

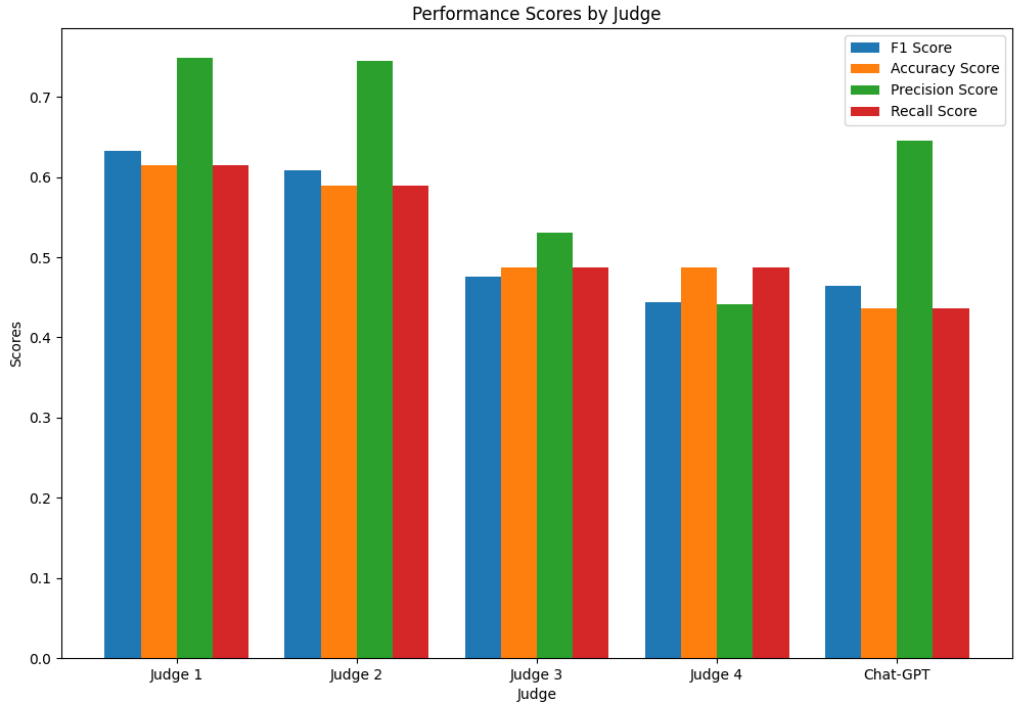

Figure 5:- The above figure shows the metric comparisions of the 4 human judges and ChatGPT compared to the gold standard/ ground truth. 1st set of bars in the bar graph show performance of judge 1 compared the the ground truth, nd set of bars show the performance of judge 2 compared to the ground truth and so on the others depict their performances.(fig above)

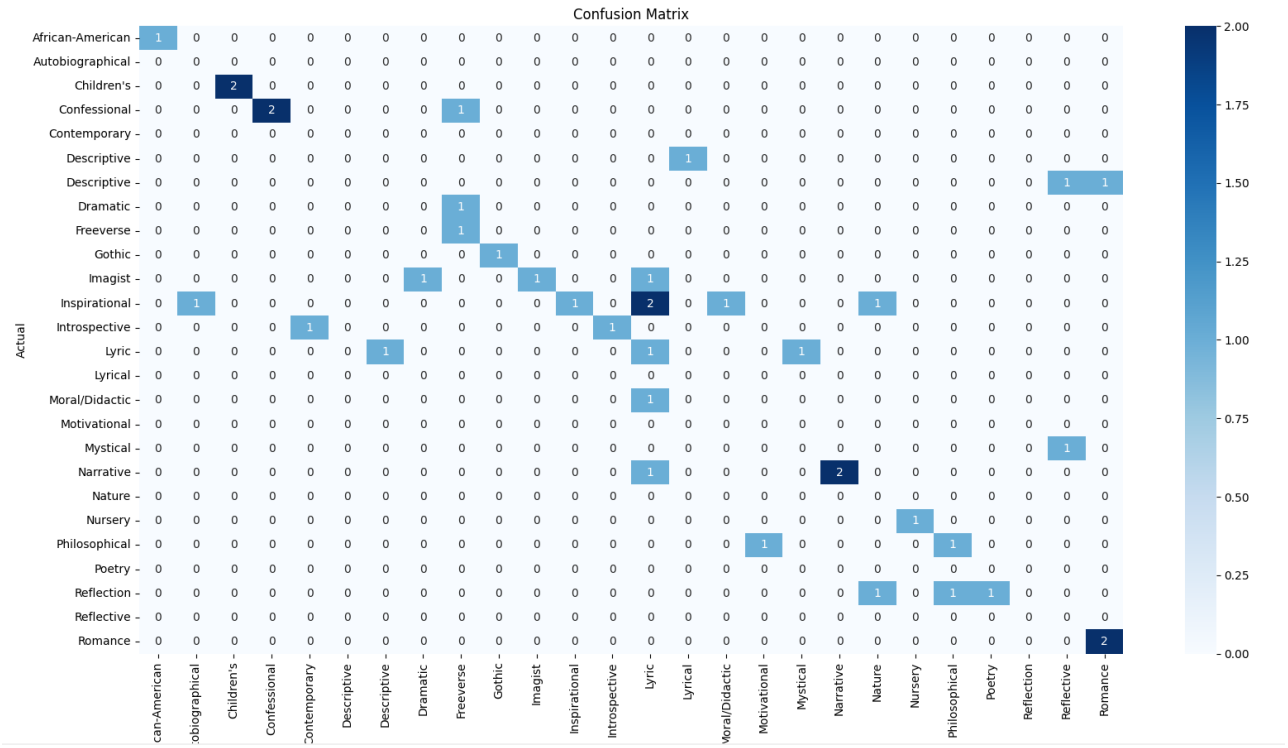

Figure 6:- This plot shows the confusion matrix of the dataset (fig above)

6.2 OBSERVATIONS

Here we clearly observe that Judge 1 has the highest F1 score with around 63 percent which means that their classifications clearly aligned with that of the majority. Then we have Judge 2 and Judge 3 ranking 2nd and 3rd respectively. Chat-GPT has ranked 4th with 46 percentage followed by Judge 4 in last Place with 44 percentage.

We see that Chat-GPT is highly "Precise" (Ranking 3rd) as compared to human behavior i.e. it shows the model's accuracy when it guesses an instance to be Positive. This precision score is balanced out by the Recall Score (Ranking 5th) , which shows the measure of how good the model is at capturing or finding all the Positive Instances.

Hence, this shows how GPT's F1-Score, which is a metric that combines the aspects of both Precision and Recall, ends up balancing itself at around the 4th Highest in our Judges. These scores hence give a valuable and comprehensive insight into the models ability to classify these poems as compared to our Human Judges.

CONCLUSION

Although Chat-GPT falls behind a slight bit as compared to human observation, it still boasts an impressive score which shows it is nothing short of an average Human when it comes to classifying these poems. Ultimately every human's opinion on anything is purely subjective to their own experience and expertise. The GPT model has revolutionized the way the World works, and with these observations we can clearly understand how it has the ability to agree to the majority of human opinion. We notice how aspects like "DAN" and "Temperature" help the model to achieve such outputs. To conclude, this goes to show the model just represents the tip of the iceberg when it comes to truly creating a sense of "Human Mimicry" in all aspects of our lives, and how achieving such a task is not far off into the future.

REFERENCES

A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models 1

A Paper by By -Junjie Ye, Xuanting Chen , Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui , Qi Zhang, Xuanjing Huang

[1] **An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges**

A Paper by - Abid Haleem , Mohd Javaid , Ravi Pratap Singh

[2] **Text analysis of ChatGPT as a tool for academic progress or exploitation**

A Paper by - Umar Ali Bukar , Md Shohel Sayeed, Siti Fatimah Abdul Razak , Sumendra Yogarayan , Oluwatosin Ahmed Amodu

[3] **Attention Is All You Need**

A Paper by -Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin

[4] **How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection**

A Paper by - Biyang Guo, Xin Zhang , Ziyuan Wang , Minqi Jiang , Jinran Nie, Yuxuan Ding , Jianwei Yue , Yupeng Wu

[5] **Simplifying the Visualizationof Confusion Matrix**

A Paper by - Emma Beauxis-Aussalet , Lynda Hardman

[6] **OpenAI's ChatGPT documentation**

Official documentation of ChatGPT from its creater OpenAI.

[7]**On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective**

A Paper by - Jindong Wang , Xixu Hu , Wenxin Hou , Hao Chen , Runkai Zheng , Yidong Wang , Linyi Yang , Wei Ye , Haojun Huang , Xiubo Geng , Binxing Jiao , Yue Zhang , Xing Xie