A Project Report

On

Understanding and critique on ChatGPT

BY

Muvvala N V Sai Gurudeep (SE20UARI103)

Sreevalli Yelisetti (SE20UARI148)

Rama Sai Pavan Kotipalli (SE20UARI123)

Under the supervision of

Assistant Prof. ARUN AVINASH CHAUHAN

ECOLE CENTRALE SCHOOL OF ENGINEERING

HYDERABAD

(May 2023)

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to all those who have contributed to the successful completion of this project. First and foremost, we extend our heartfelt appreciation to OpenAI for developing ChatGPT and making it available for research purposes, enabling us to conduct this study.

We would like to thank our project supervisor, Professor Arun Avinash Chauhan, for his guidance, support, and valuable insights throughout the project. Their expertise and encouragement have been instrumental in shaping the direction and scope of our research.

We would also like to acknowledge the efforts of the annotators who participated in this project, dedicating their time and expertise to provide valuable input and annotations. Their contributions have greatly enhanced the reliability and validity of our findings.

Furthermore, we extend our appreciation to the individuals who participated in the testing and evaluation of ChatGPT's responses, providing valuable feedback and insights. Their involvement has significantly enriched the quality and depth of our analysis.

Finally, we express our gratitude to the academic community and the broader research community for their continuous efforts in advancing the field of natural language processing and artificial intelligence. Their groundbreaking work serves as a foundation for projects like ours, driving innovation and pushing the boundaries of human-machine interaction.

We are grateful for the collective support and contributions that have made this project possible. While any errors or omissions remain our responsibility, the guidance and collaboration we received have undoubtedly played a significant role in the success of this endeavor.

ABSTRACT

ChatGPT, an advanced language model developed by OpenAI, has gained significant attention for its wide range of applications, ranging from generating text to evaluating assignments and even mimicking human-like conversations. However, amidst its remarkable capabilities, concerns have been raised about the potential ethical implications and limitations of such technology. This project delves into the critical examination of ChatGPT, aiming to understand its inner workings beyond the surface-level applications it offers.

By exploring the concept of the Turing test and the extent to which ChatGPT can mimic human behavior, the project seeks to investigate whether the machine's ability to bluff or lie challenges the traditional notion of reliability and predictability associated with computers. In this context, the project aims to deconstruct the architecture of ChatGPT, analyze its underlying mechanisms by studying the outputs, and ponder the profound questions it raises about the boundaries between machine intelligence and human consciousness.

To shed light on the issue of trustworthiness and to understand how human chatGPT is, we devised classification based testing (how well does chatGPT identify/ classify when compared to humans)This test aims to verify the authenticity of ChatGPT's responses in a manner similar to how humans interact and establish trust with one another. Additionally, since the GPT-3 architecture is not openly accessible, the project will utilize the outputs generated by ChatGPT as a basis for deconstructing its functionality, with a particular emphasis on evaluating trustworthiness.

The problem statement of the project focuses on critiquing ChatGPT and addressing the challenge of identifying the needs of individuals experiencing depression. This entails exploring whether ChatGPT can accurately recognize signs of depression and provide appropriate support. To evaluate ChatGPT's trustworthiness in this context, a comparative analysis will be conducted between human annotators and the model's classifications of depressive posts.

By critically examining ChatGPT, exploring its ethical implications, and assessing its performance in identifying depression-related needs, this project contributes to the broader understanding of human-machine interaction and the potential impact of advanced language models on mental health support systems.

CONTENTS

Introduction

Purpose

The main objective of the project is to address the challenge of accurately identifying the needs of individuals experiencing depression. It highlights the use of advanced language models like ChatGPT as potential tools for providing support to individuals with mental health issues. The subsection emphasizes the importance of evaluating the trustworthiness and effectiveness of these models in understanding and addressing the complex needs of depressed individuals. In carrying out its purpose, the project also plans to critique chatGPT - in terms of trustworthiness and 'human-ness' when it comes to language comprehension.

Project Focus

The project focuses on critiquing ChatGPT's capabilities in recognizing signs of depression and providing appropriate support. By conducting a comparative analysis between human annotators and ChatGPT's classifications of depressive posts, we aim to assess the model's performance and determine its reliability in identifying and understanding the needs of individuals experiencing depression.

Real World Application

The real-world application of this problem is significant. Depression is a prevalent mental health condition affecting millions of individuals worldwide. However, due to various reasons such as stigma, limited access to mental health resources, or reluctance to seek help, many individuals do not receive the support they need. Virtual assistants and chatbots have the potential to bridge this gap by offering accessible, anonymous, and immediate support to those in need.

By evaluating ChatGPT's ability to recognize depression-related needs, we contribute to the broader understanding of human-machine interaction in mental health support systems. This project aims to shed light on the ethical implications associated with using advanced language models for mental health support and explore their potential benefits and limitations. The findings will provide valuable insights into the effectiveness and reliability of ChatGPT and similar models in assisting individuals with depression, ultimately paving the way for the development of more accurate and empathetic virtual support systems.

Overall, this project aims to address the pressing need for reliable and effective mental health support by critically examining ChatGPT's performance in identifying depression-related needs. By doing so, we strive to contribute to the advancement of technology-driven mental health solutions and improve the well-being of individuals experiencing depression.

Problem Definition

The project aims to explore the capabilities and limitations of ChatGPT, a language model based on GPT-3 architecture. It seeks to answer several questions regarding ChatGPT's performance and its ability to understand and respond to user prompts in a meaningful and accurate manner. Additionally, the project aims to investigate the differences between GPT-3 and ChatGPT-3 and how ChatGPT addresses some limitations of GPT-3. The project also focuses to what extent can we trust ChatGPT in correctly identifying and addressing the unsatisfied needs of a depressed person, and what factors influence its reliability and accuracy in this regard? Our process begins with annotating the posts to determine their true labels. Subsequently, we engage ChatGPT to generate responses, predicting various classes and providing associated probability scores. By leveraging these probability scores for each predicted sample, we gain insights into the confidence levels, indirectly enabling us to perform a Turing test based on the classification and probability scores.

The following questions will be addressed in this project:

1. How can prompt and randomness be effectively utilized to enhance ChatGPT's performance, considering the limitations identified in GPT-3?
2. What are the specific areas where ChatGPT performs poorly, and what are the factors contributing to these limitations?
3. What are the strengths and areas of expertise where ChatGPT excels, and how does it demonstrate its capabilities effectively?
4. What are the significant differences between GPT-3 and ChatGPT-3 in terms of their functionalities and performance?
5. How does ChatGPT overcome the limitations of GPT-3, and what specific challenges or issues does it address?
6. What are the suitable temperature values for controlling the randomness of ChatGPT's responses, and how does the choice of prompt impact its output?
7. Can a set of posts be classified into different Maslow Classes, and can ChatGPT design a prompt that ranks these classes with probabilities and provides explanations?
8. How does the number of words in the window for checking probabilities influence the order of the words in ChatGPT's responses?
9. Can ChatGPT predict the unsatisfied needs of a depressed person accurately, considering its understanding of language patterns and context?
10. To what extent can we trust ChatGPT in correctly identifying and addressing the unsatisfied needs of a depressed person, and what factors influence its reliability and accuracy in this regard?

Mathematical representation is not applicable in this context as the questions primarily focus on the understanding and capabilities of ChatGPT rather than mathematical modeling or formal representations.

Background and Related Work

Our project is closely related to several notable publications in the field, which provide insights into the capabilities and limitations of GPT-3.

The research paper titled "Can GPT-3 Pass a Writer's Turing Test?"[3] examines the performance of GPT-3 (the precursor to ChatGPT) in the context of a Turing test specifically designed for evaluating its writing capabilities. The paper explores the limitations and shortcomings of GPT-3 in meeting the criteria of passing a writer's Turing test.

The research paper investigates the ability of GPT-3 to generate coherent and contextually appropriate written responses in various domains, such as storytelling, news articles, and technical writing. It aims to determine whether GPT-3 can produce outputs that are indistinguishable from those of human writers, thus passing the writer's Turing test.

The authors of the paper conduct several experiments and evaluations to assess GPT-3's performance. They employ human judges who evaluate and compare written responses generated by GPT-3 with those written by humans without knowledge of their origin. The judges rate the responses based on factors such as coherence, grammaticality, and overall quality. The paper presents the results of these evaluations and analyzes the strengths and weaknesses of GPT-3 in meeting the criteria of passing the writer's Turing test.

One of the key shortcomings identified in the research paper is the lack of deep understanding exhibited by GPT-3 in generating written responses. While GPT-3 can produce impressive outputs that often appear human-like, it often lacks semantic understanding and can generate responses that are contextually inconsistent or factually incorrect. The paper highlights instances where GPT-3 generates plausible-sounding but factually inaccurate information, demonstrating the limitations of the model's knowledge representation and reasoning abilities.

Another limitation observed in the research paper is the potential for GPT-3 to produce responses that are excessively verbose or tangential. The model may exhibit a tendency to generate lengthy and convoluted explanations, which can diminish the overall coherence and effectiveness of the written response.

Continuing the research paper titled "GPT-3: What is it Good For?"[4] further explores the strengths and limitations of GPT-3, focusing on its practical applications and areas where the model excels.

The paper begins by acknowledging the impressive language generation capabilities of GPT-3, which allows it to generate coherent and contextually relevant text across various domains. It highlights GPT-3's ability to perform tasks such as text completion, summarization, and even creative writing. However, the paper aims to delve deeper into understanding the specific contexts in which GPT-3 performs exceptionally well and where its limitations become apparent.

To achieve this, the authors conduct a series of experiments and evaluations to assess GPT-3's performance in different scenarios. They explore domains such as natural language understanding, question-answering, dialogue generation, and more. Through these evaluations, they aim to identify the specific tasks and contexts where GPT-3 demonstrates superior performance compared to other models and traditional approaches.

The research paper also discusses the limitations of GPT-3 that emerge during the evaluations. One significant limitation is the model's lack of factual accuracy and the potential for generating plausible-sounding but incorrect information. This highlights the need for caution when relying solely on GPT-3 for applications requiring precise and accurate information, such as medical advice or legal interpretations.

Furthermore, the paper addresses the issue of fine-tuning GPT-3 for specific tasks. While GPT-3 is a powerful language model, it often requires substantial fine-tuning to achieve optimal performance on specific tasks. Fine-tuning involves training GPT-3 on task-specific data, which can be time-consuming and resource-intensive. The paper discusses the trade-offs involved in fine-tuning and highlights the challenges associated with achieving a balance between retaining the general language understanding capabilities of GPT-3 while optimizing it for specific tasks.

By including this research paper in the "Background and Related Work" section of your project report, you can highlight the work done in evaluating the practical applications of GPT-3 and its performance in various domains. Additionally, you can discuss the findings regarding GPT-3's strengths and weaknesses, which will contribute to the understanding of the capabilities and limitations of similar language models like ChatGPT.

In the research paper titled "Five Ideas or How Professors Can Deal with GPT-3... For Now,"[2] the authors explore practical strategies for professors and educators to effectively incorporate GPT-3 into their teaching practices. Recognizing the potential of GPT-3 as a powerful language model, the paper aims to provide insights and recommendations on how professors can navigate the challenges and leverage the capabilities of this technology in the educational setting.

The paper begins by acknowledging the transformative impact of GPT-3 on various domains, including education. It highlights the potential of GPT-3 in automating certain teaching tasks, generating personalized feedback, and supporting student engagement. However, the authors emphasize the need for a cautious approach and offer five key ideas for professors to consider when integrating GPT-3 into their teaching practices.

Firstly, the paper suggests adopting a collaborative approach between professors and GPT-3. Rather than perceiving GPT-3 as a replacement for professors, the authors propose using the technology as a collaborative tool to enhance teaching and learning experiences. This approach emphasizes the importance of maintaining human expertise and guidance while leveraging the strengths of GPT-3.

Secondly, the research paper addresses the issue of bias in GPT-3's responses. Given that GPT-3 learns from vast amounts of data, including potentially biased sources, the authors emphasize the need for professors to critically evaluate and contextualize the outputs generated by the model. They recommend providing students with the necessary tools and knowledge to critically analyze the information presented by GPT-3.

The third idea proposed in the paper is the importance of iterative feedback and improvement. As GPT-3 is a language model, it can benefit from continuous feedback from professors and students. The authors suggest an ongoing feedback loop where professors evaluate and provide corrective feedback to the model's outputs, allowing GPT-3 to improve over time.

Additionally, the research paper highlights the significance of setting clear expectations and boundaries when utilizing GPT-3. Professors should establish guidelines and communicate to students the specific roles and limitations of GPT-3 in their educational context. This ensures that students understand the purpose of GPT-3 and its intended use as a supplementary tool.

Lastly, the paper emphasizes the need for professors to stay informed and up-to-date with advancements in GPT-3 and related technologies. They recommend actively participating in the research community, attending conferences, and engaging in discussions to understand the latest developments, ethical considerations, and best practices for incorporating GPT-3 into teaching practices.

The research paper titled "GPT-3: Its Nature, Scope, Limits, and Consequences"[5] provides a comprehensive exploration of the characteristics, scope, limitations, and potential consequences of GPT-3. The paper begins by discussing the nature of GPT-3 as a state-of-the-art language model, emphasizing its immense size and the training process involving massive amounts of text data. It highlights the model's ability to generate human-like text and its potential to revolutionize various fields such as natural language processing and content generation.

The paper then delves into the scope of GPT-3, discussing its wide range of applications across industries, including content generation, translation, chatbots, and virtual assistants. It emphasizes the transformative potential of GPT-3 in automating various language-related tasks and augmenting human creativity and productivity.

However, the paper also sheds light on the limits of GPT-3. It discusses the challenges associated with fine-tuning the model for specific tasks and the potential biases present in the training data that can influence the generated text. The authors analyze cases where GPT-3 may produce misleading or inappropriate outputs, highlighting the importance of cautious and responsible use of the model in sensitive applications.

Furthermore, the research paper explores the consequences of widespread adoption and reliance on GPT-3. It discusses the ethical considerations associated with the model's capabilities, including issues related to privacy, data security, and the potential impact on human employment and social dynamics. The authors emphasize the need for robust guidelines and regulations to ensure responsible and ethical deployment of GPT-3 and similar language models.

The research paper titled "Modeling Naive Psychology of Characters in Simple Commonsense Stories"[6] focuses on the modeling of naive psychology within the context of simple commonsense stories. The paper explores the understanding and representation of characters' mental states, intentions, beliefs, and desires, which are essential aspects of human cognition and reasoning.

The authors propose a computational model that aims to capture the nuances of naive psychology by utilizing a combination of language representation techniques, knowledge bases, and commonsense reasoning. The model is designed to interpret and generate narratives, allowing for the analysis of characters' mental states and the prediction of their behaviors based on their beliefs and desires.

Through a series of experiments and evaluations, the researchers assess the model's performance in understanding and reasoning about characters' mental states. They examine how well the model can predict character actions, anticipate their goals and intentions, and infer their beliefs and desires from the narrative context.

However, the paper also discusses the challenges and limitations associated with modeling naive psychology. One of the key challenges is the ambiguity and subtlety often present in character motivations and intentions. It is challenging to capture the complexity and diversity of human behavior accurately, especially in scenarios where characters' beliefs and desires might not align with their actions.

Additionally, the paper acknowledges the limitations of relying solely on text-based data for modeling naive psychology. While textual narratives provide valuable insights, they may lack real-time contextual information and nonverbal cues that play a crucial role in understanding human behavior. Incorporating multimodal data, such as audio or visual cues, could potentially enhance the accuracy and richness of the model's representation of characters' mental states.

By analyzing and synthesizing the insights from various research papers, this project gains a comprehensive understanding of ChatGPT, an AI language model based on GPT-3 architecture, and its implications for the problem statement. The papers provide valuable insights into the capabilities and limitations of GPT-3 and its practical applications. While GPT-3 demonstrates impressive language generation abilities, it falls short of passing a writer's Turing test, indicating the need for improvement. Additionally, the papers shed light on the importance of fine-tuning GPT-3 for specific tasks and addressing its limitations in terms of factual accuracy. The project also considers the ethical implications of widespread adoption of GPT-3, emphasizing the need for responsible deployment and regulatory guidelines. Furthermore, the exploration of modeling naive psychology in narratives provides a framework for understanding human behavior, which can be applied to enhance ChatGPT's ability to identify and address the unsatisfied needs of depressed individuals. The project aims to leverage these research findings to evaluate ChatGPT's performance, assess its trustworthiness, and investigate factors influencing its reliability in addressing the needs of depressed individuals.

Implementation

The implementation strategy for this project focuses on evaluating the capabilities and limitations of ChatGPT. The project aims to address the problem of identifying and understanding the needs of individuals experiencing depression. The implementation strategy can be divided into several key steps as given below:

Parameter Experimentation and Prompt Engineering

- The first step involves experimenting with different parameters of ChatGPT to identify the optimal settings for obtaining responses efficiently.
- Additionally, understanding prompt engineering techniques is crucial to prompt ChatGPT effectively for desired responses.
- By fixing the appropriate parameters and framing suitable prompts, we lay the foundation for the subsequent steps.

Obtaining ChatGPT Responses and Classifications

- We acquire responses from ChatGPT for a given set of posts, aiming for accurate classifications of the posts.
- ChatGPT returns classes with their corresponding probabilities and explanations, providing insights into its classification process.

Human Annotation and Majority Label

- Simultaneously, human annotators annotate the same set of posts for which ChatGPT generates responses.
- The majority label is determined by considering the annotations provided by the annotators.
- For simplicity, we consider only one label per post, though future work may explore the possibility of multiple labels.

Annotator Guidelines and Inter-Annotator Agreement

- Annotators are provided with a rule book detailing the categorization of posts based on human-machine understanding.
- The quality of annotation is assessed by evaluating the inter-annotator agreement scores between different annotators.

- If agreement scores are not satisfactory, annotators refine their annotation rules, reannotate the posts, and re-evaluate inter-annotator agreement.

Handling Indications of Needs and Depressed Posts

- Posts lacking indications of needs from individuals experiencing depression are removed from the dataset.
- These removed posts are replaced with a new set of posts to ensure a comprehensive coverage of relevant content.

Calculation of Flesch's Kappa and Cohen's Kappa Scores

- Flesch's Kappa and Cohen's Kappa scores are computed to assess the inter-annotator agreement results.
- These scores provide a measure of agreement between annotators and assist in evaluating the quality and consistency of annotations.

Comparison of True Labels and Predicted Labels

- The true labels obtained from human annotations are compared with the predicted labels generated by ChatGPT.
- This comparison allows us to evaluate the classification accuracy of the ChatGPT model.

Error Estimate and Confidence Interval

- To estimate the error of the model, a 0.95 confidence interval is calculated, providing a range within which the error percentage lies.
- This analysis helps in understanding the level of trust we can place in ChatGPT's classification of needs in depressed posts.

Future Steps

- After obtaining the results, the labels and explanations provided by ChatGPT are revealed to the annotators to observe any potential changes in their answers. Classification scores are recalculated accordingly.
- A comparison is made between the outputs of ChatGPT and other models such as a Machine Learning Classifier or a Deep Learning Language Model to understand variations and explore alternative approaches.

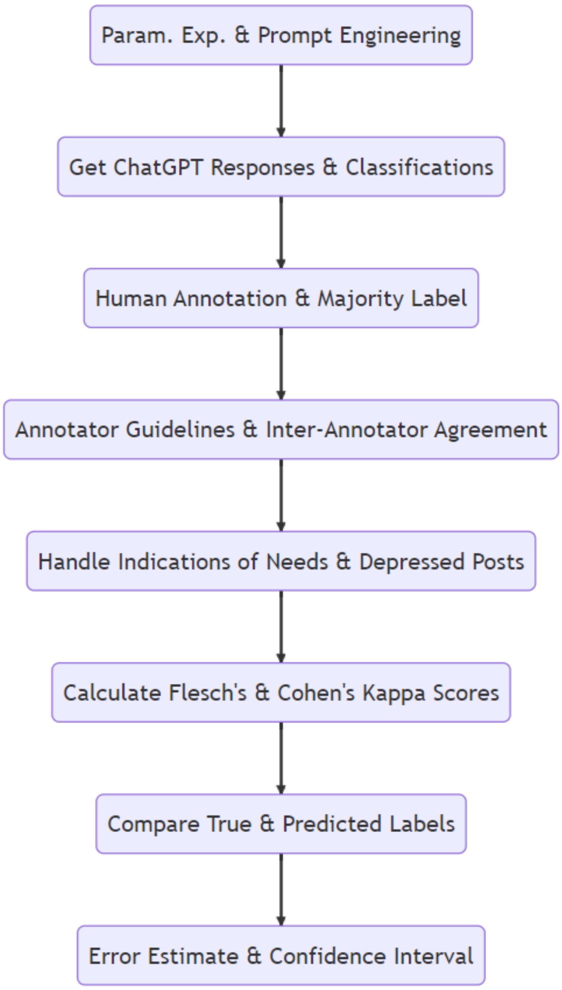

Param. Exp. & Prompt Engineering
Get ChatGPT Responses & Classifications
Human Annotation & Majority Label
Annotator Guidelines & Inter-Annotator Agreement
Handle Indications of Needs & Depressed Posts
Calculate Flesch's & Cohen's Kappa Scores
Compare True & Predicted Labels
Error Estimate & Confidence Interval

Figure 1 Implementation Strategy FlowChart

Results

Result:

The project's findings are presented in Figures 2 to 7, offering a comprehensive analysis of the performance and reliability of ChatGPT in identifying depressive needs. Figure 2 showcases the confusion matrix, which measures the agreement among annotators using the Fleiss Kappa score. The substantial agreement of 0.7435 and the consistent scores ranging from 0.716 to 0.77 indicate a high level of agreement among the annotators.

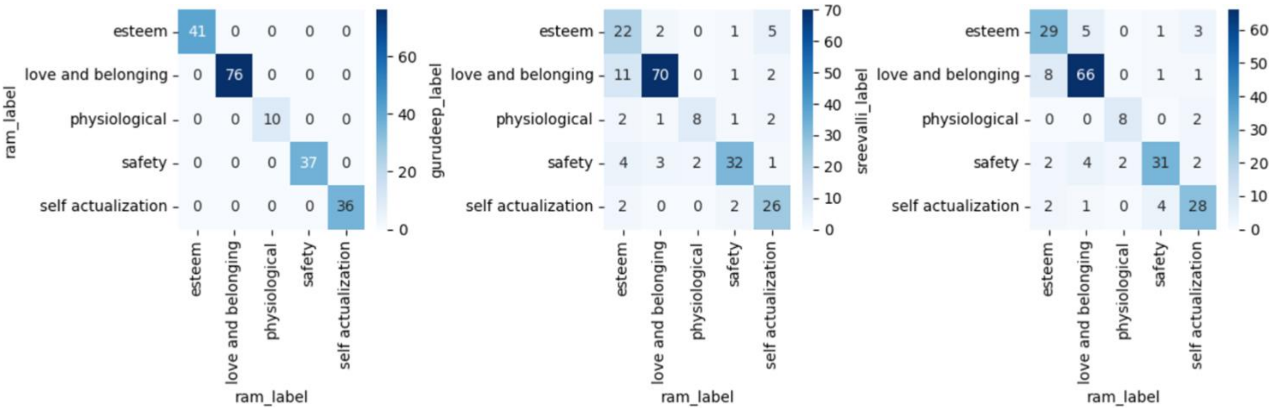
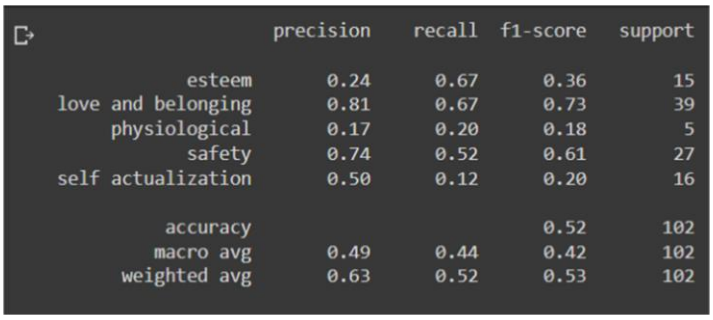


Figure 2 Confusion Matrix for Inter Annotator Agreement

Figure 3 provides the classification report for ChatGPT's classification of depressed posts into different Maslow Needs, while Figure 4 displays the confusion matrix of the model. The model demonstrates proficiency in identifying love, belongingness, and safety needs, but faces challenges with esteem and self-actualization. Additionally, the model's prediction for the physiological class performs poorly, likely due to the limited number of samples in that particular class.



```
                    precision    recall  f1-score   support

            esteem       0.24      0.67      0.36        15
  love and belonging       0.81      0.67      0.73        39
      physiological       0.17      0.20      0.18         5
             safety       0.74      0.52      0.61        27
  self actualization       0.50      0.12      0.20        16

          accuracy                           0.52       102
         macro avg       0.49      0.44      0.42       102
      weighted avg       0.63      0.52      0.53       102
```

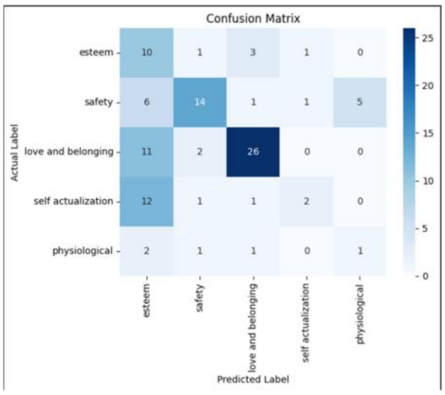


Figure 3 Classification Report of the Model

Figure 4 Confusion Matrix for the predictions

Confidence in the model's predictions is statistically computed using a Confidence Interval of 0.95, yielding a lower limit of 0.486 and an upper limit of 0.624. Figure 5 shows confidence levels for all the classes. This analysis provides insights into the model's confidence levels across all classes.

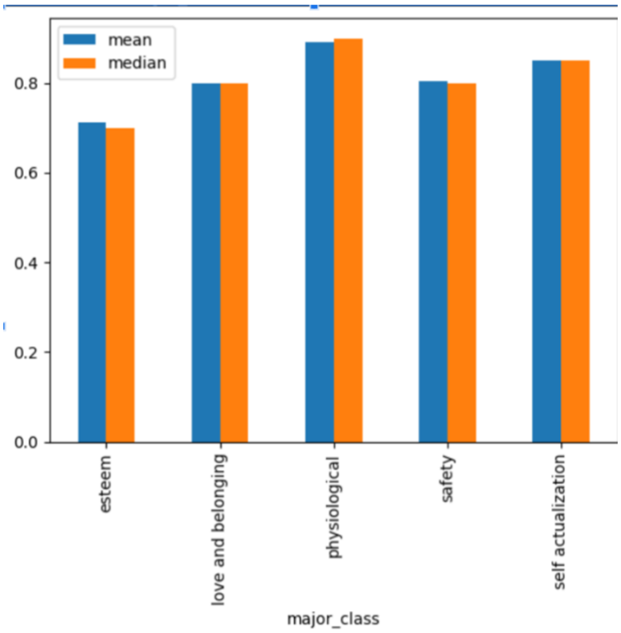


Figure 5 Confidence Levels for all the categories

To determine the optimal probability threshold for considering ChatGPT predictions on the needs of depressed users, Figure 6 presents prediction errors by probability. This analysis aids in understanding the reliability of the model's predictions and assists in decision-making.

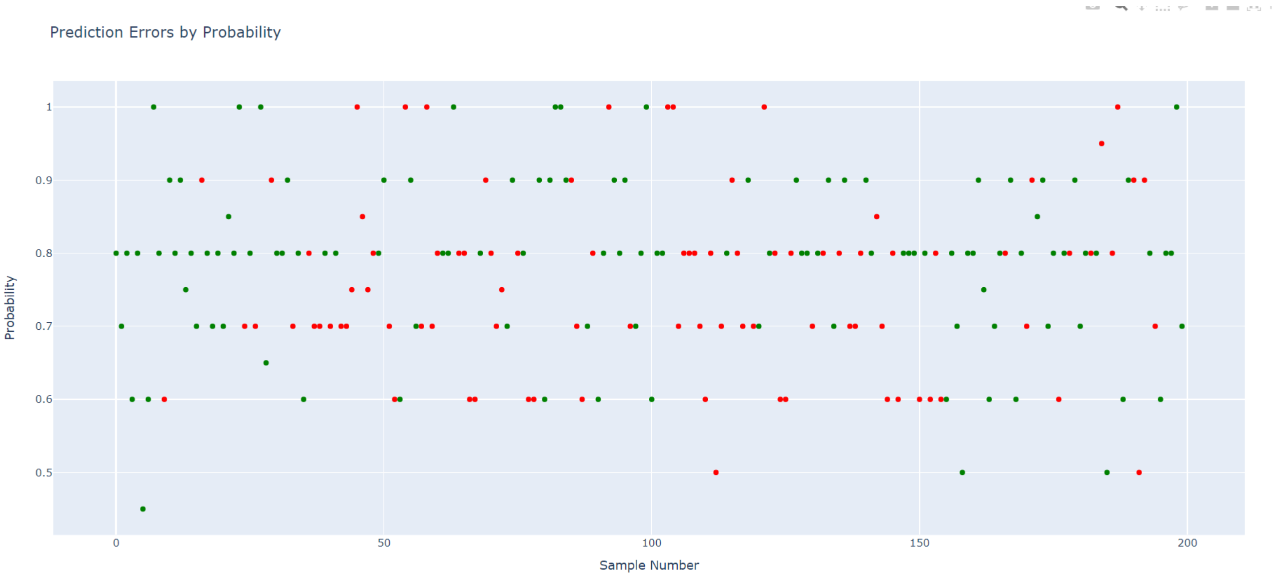


Figure 6 Confidence Levels for all the categories

Figure 7 offers an overview of the model's prediction quality on different needs, utilizing probability scores. These results highlight the model's strengths and weaknesses in identifying depressive needs, with specific attention to the areas of esteem and self-actualization.

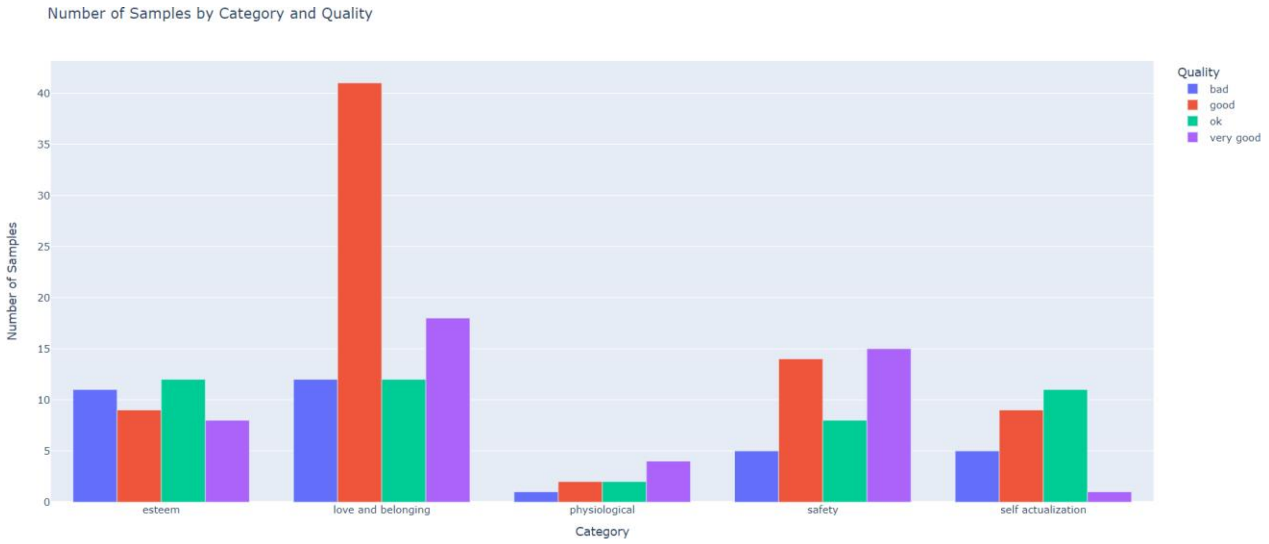


Figure 7 Category and Quality of Predictions on different needs

Overall, the results emphasize ChatGPT's performance in identifying depressive needs and provide valuable insights for potential improvements. The confidence analysis and prediction error assessment contribute to a deeper understanding of the model's reliability in this context, paving the way for future advancements in human-machine interaction and mental health support systems.

Conclusion

In conclusion, this project delved into a critical examination of ChatGPT, an advanced language model developed by OpenAI, aiming to understand its inner workings beyond its surface-level applications. The project explored the concept of the Turing test and evaluated the extent to which ChatGPT can mimic human behavior, raising questions about the reliability and predictability traditionally associated with computers. By analyzing the outputs and deconstructing the architecture of ChatGPT, the project shed light on its functionality and ethical implications.

The results presented in Figures 2 to 7 provide valuable insights into the performance of ChatGPT in identifying depressive needs. The substantial agreement among annotators, as indicated by the Fleiss Kappa score, demonstrates a high level of consistency. The model excels in identifying love, belongingness, and safety needs but struggles with esteem and self-actualization. The prediction for the physiological class is also hindered due to limited sample availability. Confidence analysis reveals the model's varying levels of certainty, and prediction error assessment helps determine the best probability threshold for considering ChatGPT predictions.

These findings contribute to a broader understanding of human-machine interaction and highlight the potential impact of advanced language models on mental health support systems. While ChatGPT shows promise in identifying depressive needs, improvements are necessary, particularly in addressing the challenges related to esteem and self-actualization. Moreover, the ethical implications of machine intelligence and its boundaries with human consciousness warrant further exploration.

This project underscores the significance of critically examining advanced language models like ChatGPT and calls for continued research and development to enhance their trustworthiness, reliability, and effectiveness in addressing complex human needs. By acknowledging the limitations and strengths of ChatGPT, we can strive towards leveraging these technologies responsibly, ensuring their positive impact on mental health support and fostering meaningful human-machine collaborations.

References

1. Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries?. Library Hi Tech News, 40(3), 26-29.
2. Pickell, T. R., & Doak, B. R. (2023). Five Ideas for How Professors Can Deal with GPT-3... For Now.
3. Elkins, K., & Chun, J. (2020). Can GPT-3 pass a Writer’s turing test?. Journal of Cultural Analytics, 5(2).
4. Dale, R. (2021). GPT-3: What’s it good for?. Natural Language Engineering, 27(1), 113-118.
5. Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694.
6. Rashkin, H., Bosselut, A., Sap, M., Knight, K., & Choi, Y. (2018). Modeling naive psychology of characters in simple commonsense stories. arXiv preprint arXiv:1805.06533.
7. Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:2303.15056.
8. https://www.theverge.com/21346343/gpt-3-explainer-openai-examples-errors-agi-potential
9. https://www.vox.com/future-perfect/21355768/gpt-3-ai-openai-turing-test-language
10. https://algowriting.medium.com/gpt-3-temperature-setting-101-41200ff0d0be
11. https://www.oreilly.com/radar/what-are-chatgpt-and-its-friends/
12. https://www.reddit.com/r/GPT3/comments/o1qvk7/are_tokens_syllables/
13. https://towardsdatascience.com/exploring-token-probabilities-as-a-means-to-filter-gpt-3s-answers-3e7dfc9ca0c
14. https://deepchecks.com/model-confidence-and-how-it-helps-model-validation/
15. https://machinelearningmastery.com/confidence-intervals-for-machine-learning/