PROJECT REPORT - DO LARGE LANGUAGE MODELS HAVE A PERSONALITY

FRANK LI (ZL3204), LINHAO YU (LY2590), MOUWEI LIN (LM3756),
XINGYE FENG (XF2248), YUTA ADACHI (YA2488)

Data Science Institute, Columbia University, New York City, NY

1. Problem Statement and Overall Approach

The evolution of Language Learning Models (LLMs) like ChatGPT represents a significant paradigm shift in human-machine interactions. These advanced LLMs not only process information and generate outputs but also simulate interactions with nuances akin to human 'personality traits'. Understanding, controlling, and potentially modifying these traits is at the heart of this research.

1.1. **Problem Statement.**

The expanding integration of LLMs across various sectors, from customer service to personalized virtual assistants, demands a thorough understanding of their behavioral patterns. This need is driven by both functional objectives and ethical considerations. For instance, an AI perceived as 'friendly' could significantly impact user trust and comfort. On the flip side, the potential of AI to reflect societal biases poses a real challenge. Consequently, this research seeks to dissect, understand, and, where possible, shape the perceived personalities of these models, balancing functional efficacy with ethical responsibility.

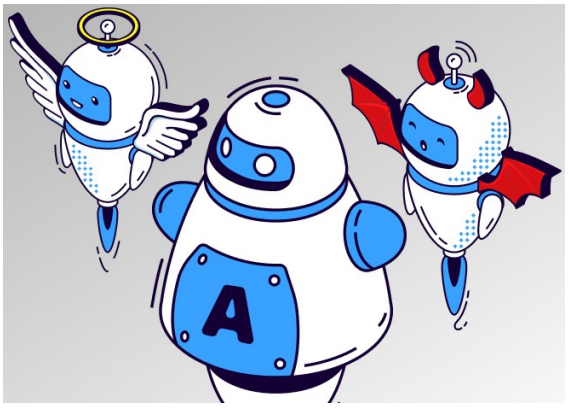

Figure 1. Dual impact of AI
Source: qulix.com

Date: December 15, 2023.

1.2. **Overall Approach.** Our research adopts a multi-faceted approach to explore the personality traits of LLMs. Initially, we utilize existing machine learning models to analyze responses generated by ChatGPT to various prompts, aiming to identify patterns indicative of distinct personality traits. This analysis is complemented by a thorough review of relevant literature, providing context and framing our empirical observations.

Next, we focus on refining our LLM into a specialized text-based personality detection model. This model is applied to assess the personality traits revealed in ChatGPT-generated tweets and comments. The comparison of results between stages one and two enables us to evaluate the consistency and reliability of detected personality traits in LLMs.

Finally, we employ a dual-method self-test, integrating both direct and indirect querying approaches, to assess the behavior of GPT 3.5. This includes adapting structured personality test prompts and transforming them into hypothetical scenarios for indirect evaluation. This comprehensive approach allows us to scrutinize the nuances of LLM personality traits, assess the influence of query types on responses, and validate the robustness of our findings.

2. Theoretical Background

The preliminary phase involves delineating the concept of 'personality' in the realms of human psychology and artificial intelligence. A comprehensive understanding of 'personality' in the context of an AI, juxtaposed with human personality metrics, serves as the foundation for this exploration.

2.1. **Background of Personality Metrics.**
Big Five Traits. Central to human psychological profiling are the Big Five personality traits. These encompass Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, collectively represented by the acronym OCEAN in Figure 2. An in-depth understanding of these traits facilitates the exploration of potential parallels in LLM responses and behaviors.
MBTI. The Myers-Briggs Type Indicator (MBTI) in Figure 3 offers an alternative perspective on personality, detailing 16 distinct types based on trait combinations, such as introversion/extraversion and thinking/feeling. Applying the MBTI framework to AI analysis can uncover novel insights into LLM behavior.

2.2. **Application to LLMs.** Applying human personality metrics to AI models like ChatGPT is a novel endeavor. The objective is to discern patterns, similarities, and divergences. Through methodical administration of prompts and subsequent analysis of responses, this research aims to draw correlations between the LLM's exhibited behaviors and recognized human personality traits.

2.3. **Background of LLaMA 2 Model.** The LLaMA (Touvron et al., 2023a) model in Figure 4 represents a significant advancement in the field of large language models. Developed as a more efficient and scalable alternative to its predecessors, LLaMA 2 (Touvron et al., 2023b) is designed to process and generate natural language with improved accuracy and context understanding. This model leverages deep learning techniques, particularly transformer architectures, to analyze and predict text sequences, making it highly effective in understanding and generating human-like responses.

The core of LLaMA 2's capability lies in its extensive training on diverse datasets, encompassing a wide range of topics and writing styles. This training enables the model to exhibit a broad understanding of language nuances, context, and even the subtleties of human emotion and thought processes. Such a comprehensive language understanding framework makes LLaMA 2 particularly suitable for personality detection tasks, as it can accurately interpret and respond to prompts that require deep language comprehension and contextual awareness.

In the context of this study, LLaMA 2's advanced language processing abilities are crucial. By employing this model for personality detection, we aim to push the boundaries of what AI can discern about human-like personality traits. The model's proficiency in handling complex

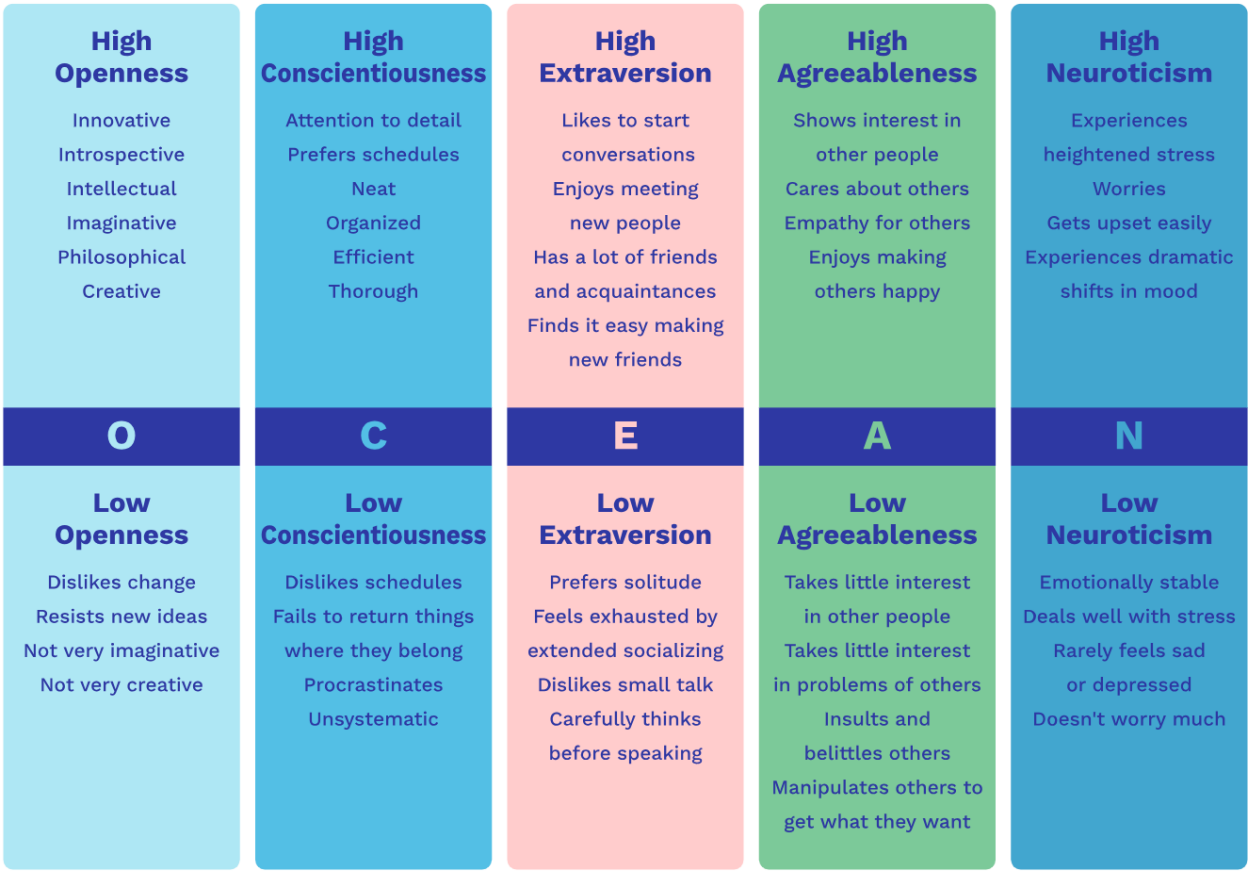

FIGURE 2. Diagram of the Big Five Personality Traits
Source: embrace-autism.com/big-five

language patterns and its ability to generate nuanced responses allows for a more detailed and accurate assessment of personality traits in large language models. This enhances our ability to draw parallels and identify divergences between AI-generated responses and human personality characteristics.

2.4. **Data Collection and Initial Observations.** In the construction of a model designed to predict MBTI types from textual input, this study utilizes a dataset sourced from Kaggle. The dataset is comprised of over 8,600 entries, each corresponding to an individual's MBTI type. Each entry includes excerpts from the last 50 posts made by the individual on the PersonalityCafe forum, which are separated by a sequence of three pipe characters ("|||").

3. METHODOLOGY

3.1. **Frameworks of Project.** As introduced, our preliminary goal for the first stage is to use an existing machine-learning model to predict LLMs' personalities by analyzing the responses they generate and briefly going over previous research on this topic. The responses we would like them to generate are tweets and essays according to the different prompts we give them. We reviewed some literature on previous works on this topic. And we referred some to start our own experiments and validation. The second stage is to fine-tune our own LLM model into a text-based personality detection model, then test the personality with the tweets and comments generated by ChatGPT, and compare the first stage's results to see which model has better performance and if ChatGPT has a personality and if yes, how consistent it is and how reliable the result is. The third stage, also the

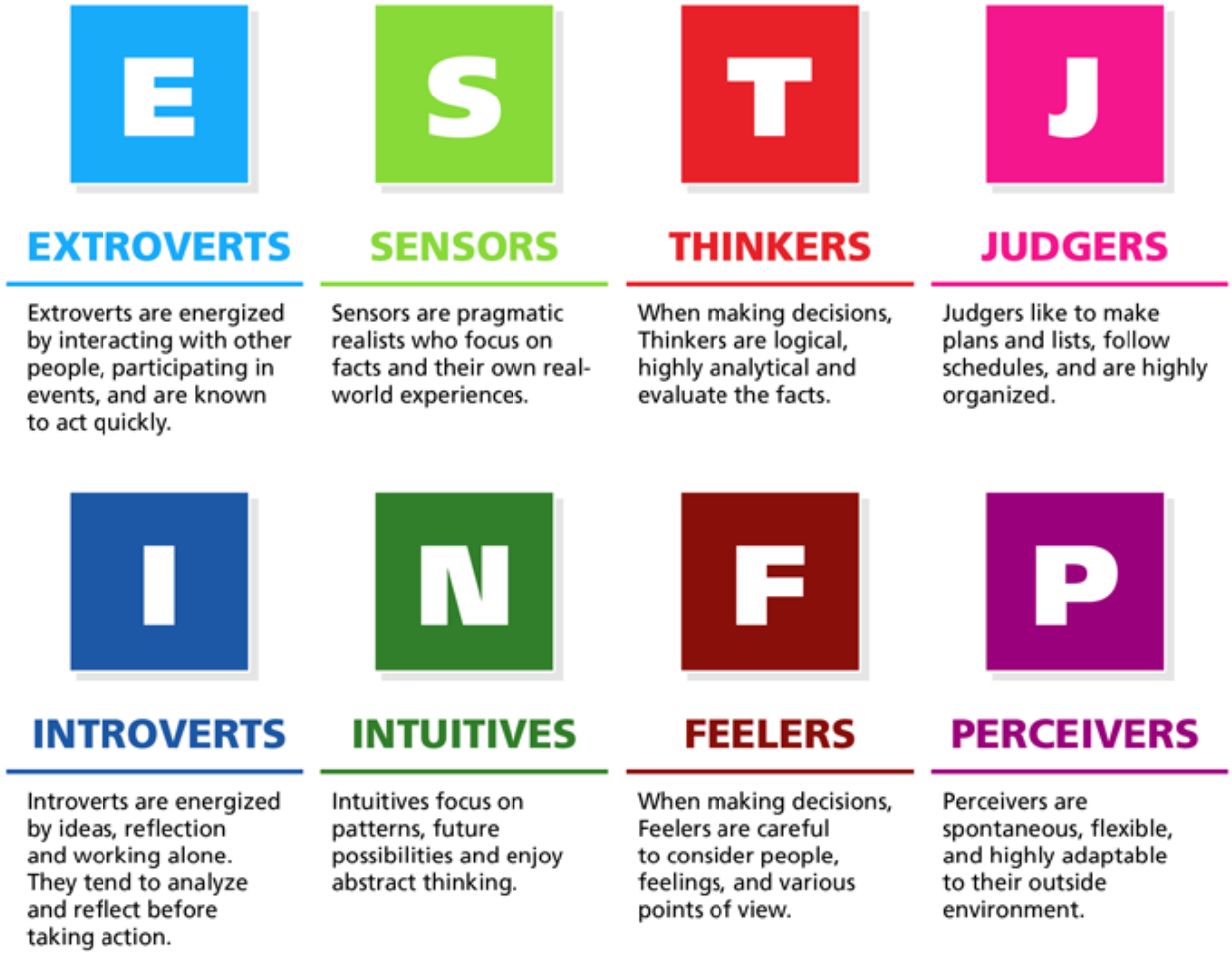

FIGURE 3. MBTI Diagram
Source: leadx.org/articles/mbti-myers-briggs-type-indicator-overview

last stage, is to do a Dual-method self-test, which obtains evaluation scores of the behavior of GPT 3.5 with two methods. This stage aims to blend direct and indirect query approaches to explore the depth and nuances of LLMs' personality traits. Here, we integrate structured prompts from recognized personality tests (OCEAN and MBTI) and transform them into scenarios to indirectly gauge ChatGPT's responses. This dual-method strategy allows us to cross-verify the consistency of ChatGPT's personality traits and explore how different querying techniques might influence its responses. The goal is to assess the robustness of LLM personality traits and to understand the extent to which ChatGPT's responses are shaped by the nature of the queries, thus contributing to a more nuanced understanding of LLMs in the context of personality detection and behavioral prediction. Figure5 is the flow chart of the last two stages.

3.2. **First Stage: Literature Review and Research Result Reproduction.** Our approach by referring to the literature, includes reviewing existing text-based personality prediction models, generating prompts and responses, and finally using the prediction models to predict LLM personality according to the prompts. We focus on GPT 3.5 as our target LLM and the MBTI test as our benchmark personality test in this experiment.

3.2.1. *Personality Detection Model.* A common flow of predicting personality using machine learning models is shown in this figure6. While defined in the experiment design above that the different

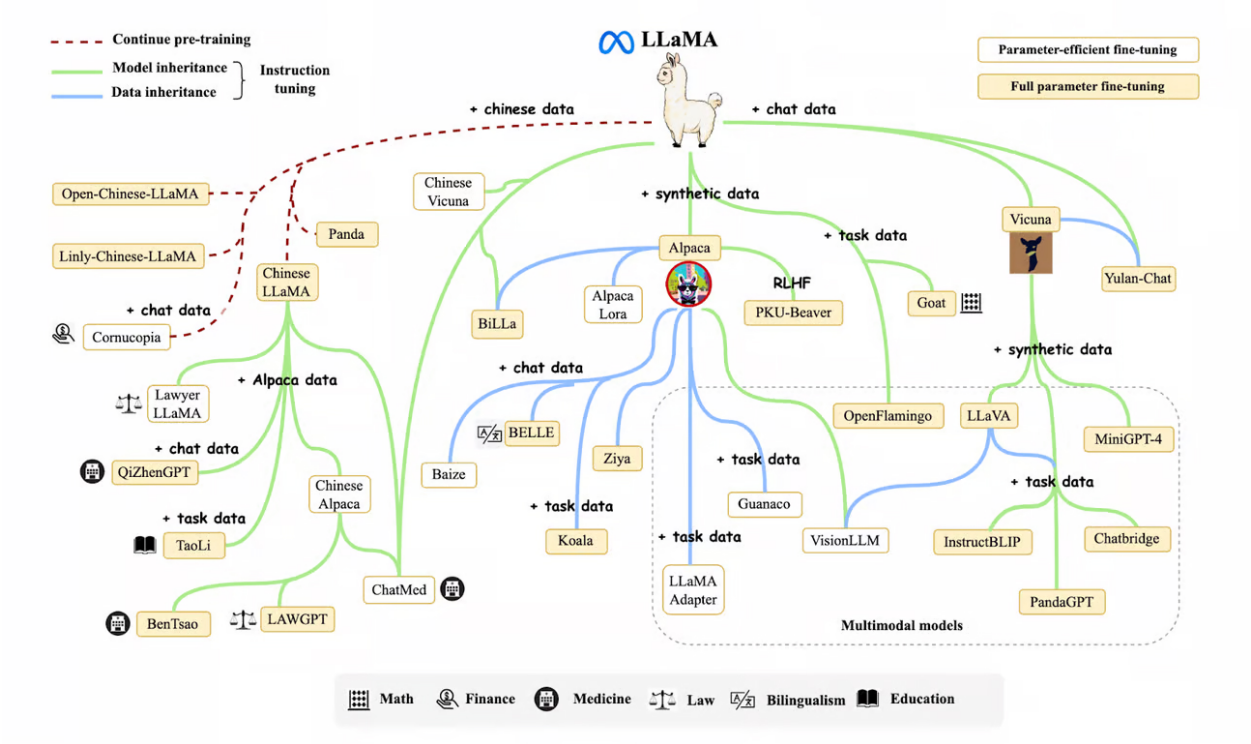

FIGURE 4. LLaMa application
Source: doi.org/10.48550/arXiv.2303.18223

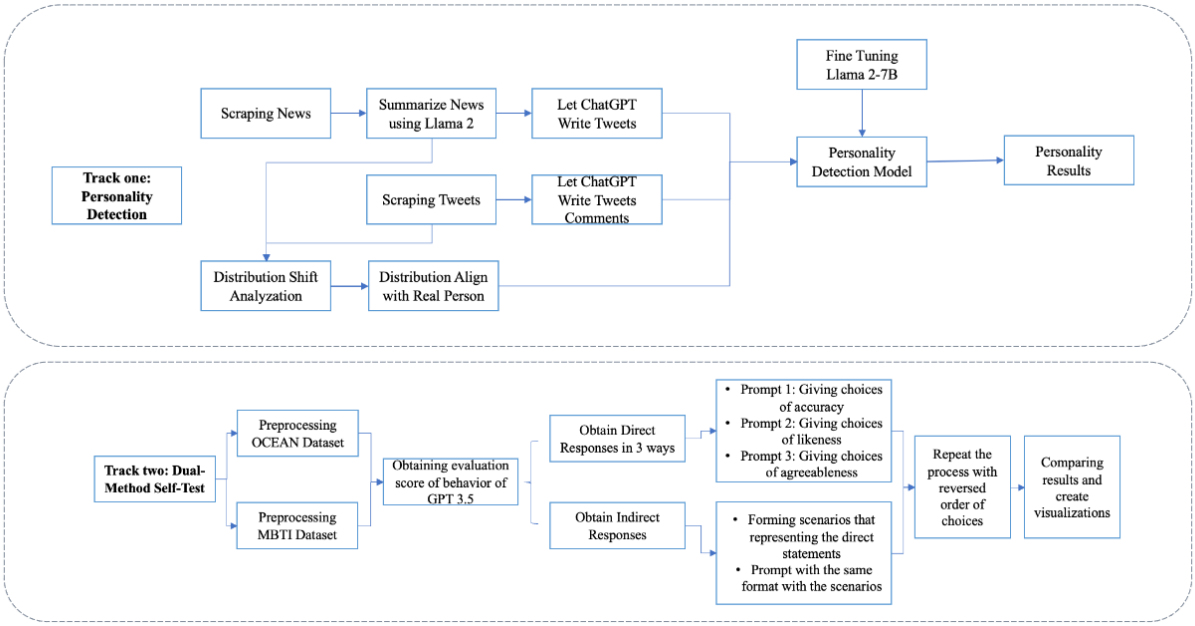

FIGURE 5. Flow Chart of Last Two Stages

text responses generated by GPT will serve as the "Social Media Posts" in the flow figure, we need to process them in a "Personality Detection Model" in order to output the "MBTI Personality Traits" predictions. We looked at three different prediction models from the literature we read.

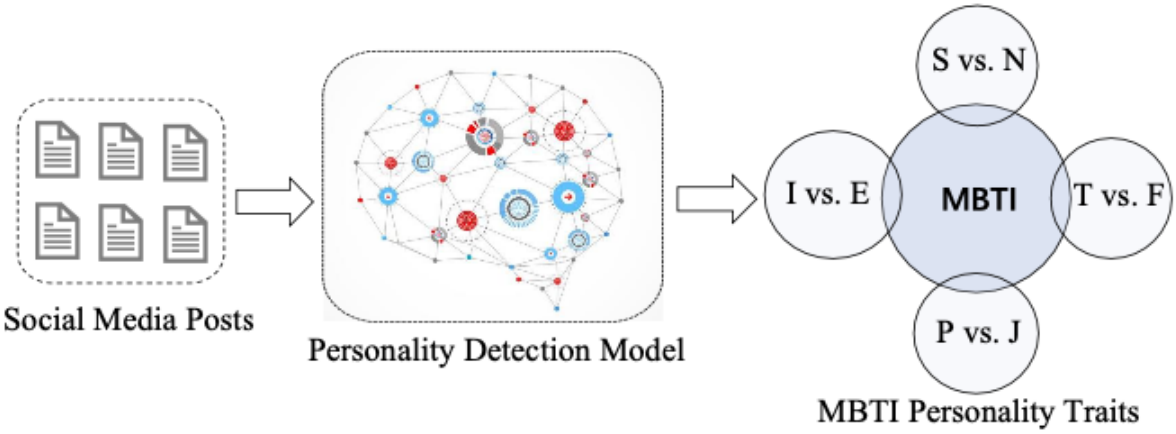

FIGURE 6. Personality Detection Model Flow

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...
...	...	...
8670	ISFP	'https://www.youtube.com/watch?v=t8edHB_h908||...
8671	ENFP	'So...if this thread already exists someplace ...
8672	INTP	'So many questions when i do these things. I ...
8673	INFP	'I am very conflicted right now when it comes ...
8674	INFP	'It has been too long since I have been on per...

FIGURE 7. Kaggle MBTI Dataset

(1) Model 1: Text processing: BERT, Prediction classifier: MLP. (Yash Mehta et al.)
(2) Model 2: Text processing: BERT, Prediction classifier: GNN. (Tao Yang et al.)
(3) Model 3: Text processing: RoBERTa, Prediction classifier: attention-based denoising framework (ADF). (Qirui Tang et al.)

All models were trained on a Kaggle dataset of social media posts that have the MBTI trait labeled for each person like figure 7. Then used BERT/RoBERTa for feature extraction, and went through their respective prediction classifiers to predict the MBTI traits.

3.2.2. *Prompt Generation and Testing.* Following the experiment design to generate prompts for GPT to answer, we then used a ChatGPT API to embed GPT answers in our Python coding environment. So the answers are ready to be processed and be used in the classification model to predict MBTI traits. The prediction output is a possibility of classifying to traits "E"(as opposed

to "I"), "S"(as opposed to "N"), "T"(as opposed to "F"), "J"(as opposed to "P"), in a scale of 0 - 1. Details of the prompts we used in the experiment are attached in Appendix I.

3.3. **Second Stage: ChatGPT Personality Detection.** We fine-tune a LLaMA 2(Touvrone-tal.,2023) into a text-based personality detection model, then let ChatGPT generate tweets and comments based on some topics, and news we scrap from the Internet. Finally we do some topic distribution analysis and personality result reliability test.

3.3.1. *News and Comments ETL Pipeline.* The two text formats we want ChatGPT to generate are tweets and comments after reading real-world news, articles and tweets. To scrape large amounts of daily news fast and efficiently, we designed an ETL pipeline in Python, using newspaper API to catch news, and deployed it with multi-thread to speed up the process, then formalize the format. However, since each article usually has many words, we use another Llama2-7B to summarize each news into 200-300 words before prompting into ChatGPT. Once we had summarized news, we did some prompt engineering to let ChatGPT read and write a tweet about the news we just put in.

To generate ChatGPT comments in response to commonly posted tweet categories, we scraped 5000 tweets which contain 10 topics: Bitcoin, NFL, Music, Oscars, Travel, Fashion, Food, Fitness, Gaming, Art, Technology, and ThrowbackThursday nearly evenly distributed. The topics were chosen by asking ChatGPT to output the Top 20 Twitter topics of all time and randomly picking 10 from them. Then we asked ChatGPT to write a comment in response to each scraped tweet. The reason for generating the comments is to mimic users' posts on Twitter which usually consist of initial tweets written by users and also reposts and comments to others' Tweets.

After collecting enough generated tweets, we would like to see the topic distribution of those generated tweets compared with the real-world tweets we used to train our own personality detection model. We randomly selected 5 thousand generated tweets and 10 thousand training tweets and put them into our Llama2-7B summarizing model to generate classification labels, we found that topic distribution of training tweets vs. testing generated tweets from the news is significantly shifted, which means it is very interesting to see what will be the distribution of personality under such shift. If the distribution is still consistent and stable, it will be strong evidence to tell us that LLM, at least ChatGPT, has a personality. Figure8

3.3.2. *Fine-tuning Personality Detection Model.* We fine-tuned the LLaMA 2 7B (Touvron et al., 2023b) model using Low-Rank Adaptation (LoRA) (Hu et al., 2022) to enhance personality detection capabilities based on the MBTI. LoRA introduces trainable low-rank matrices into specific layers of a pre-trained model, enabling efficient fine-tuning while preserving the model's structure. We developed four distinct binary classification models, each for one of the MBTI dimensions, using a dataset exclusively from Kaggle, divided into training, evaluation, and testing subsets in an 81:9:10 ratio.

The fine-tuning process was conducted with a learning rate of 1e-4 over 5 epochs, and a global batch size of 8. The training started with a warm-up phase for the first 100 steps, followed by a linear decay in the learning rate.

Our methodology for optimizing LoRA settings involved a multi-step approach, focusing on the rank parameter r and the target modules. We initially trained models with target modules set to the query (q) and value (v) layers, testing r values of 8 and 16. The r value that led to higher evaluation accuracy was then used in subsequent training with all linear layers as target modules. The primary metric for performance evaluation was accuracy, guiding our hyperparameter settings. The final model selection for each MBTI dimension was based on the highest evaluation accuracy, ensuring the most effective approach for MBTI personality detection.

3.3.3. *Personality Detection.* We packed generated tweets or comments into a format that exactly the same as training samples. Which is, for each training sample, there will be 50 tweets written by

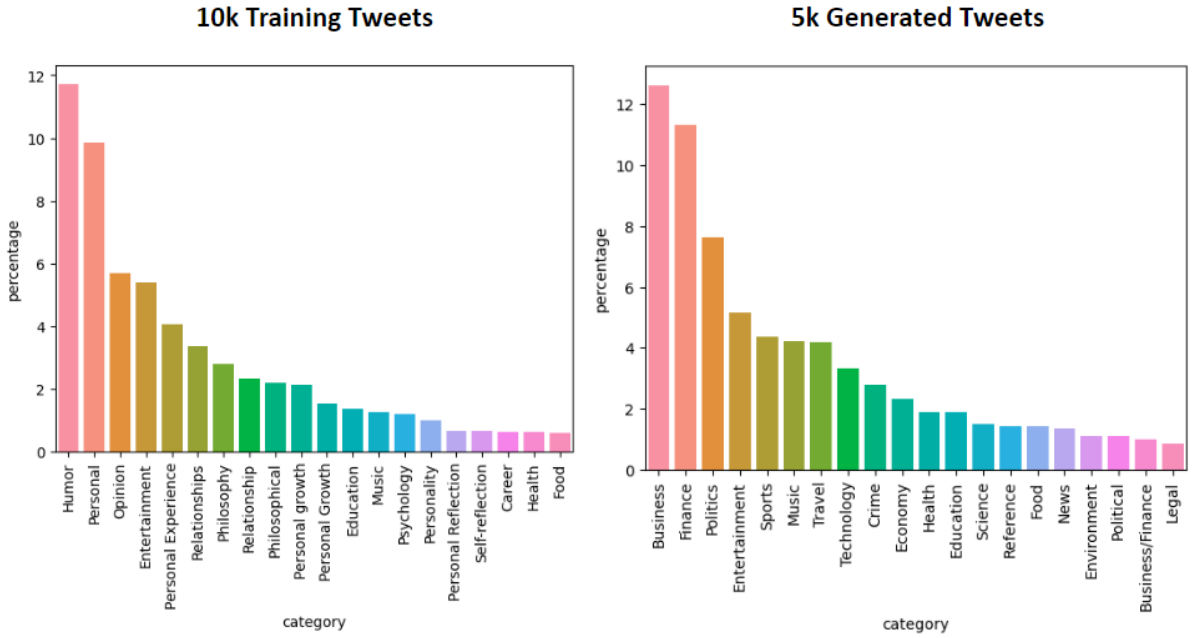

FIGURE 8. Topic Distribution of Training Tweets vs. Testing Generated Tweets from News. Topic shift noticed.

the same person concatenated together and separated by number. So we do the same transformation, randomly selecting 50 tweets or comments with or without replacement. Then combined them as one testing sample, and after that, we put it into our model, it will finally output one MBTI label which has the highest probability among a total of 16 labels. For example, input sample: """1. A tweet 2. A tweet .... 50.A tweet """, then the output might be INFJ.

3.3.4. *Self-Assessment Test.* In the final phase of the second stage, we conducted a self-assessment test to evaluate the validity and consistency of ChatGPT's MBTI results as determined by the personality detection model. For this assessment, we utilized a questionnaire dataset from Pan and Yawen (2023) that was originally in Chinese and translated into English. The questionnaire consisted of 93 items, each posing a binary choice that reflects distinct personality traits. For example, a typical item might ask: "When you are going out for the whole day, do you: A. Plan what to do and when to do it, or B. Not make plans, and just go with the flow?" Each choice, A or B, corresponds to specific MBTI dimensions such as Judging (J) or Perceiving (P).

The test was administered to ChatGPT, which involved presenting each question and recording ChatGPT's selection between options A and B. Following the completion of the questionnaire, the responses were aggregated, and the most frequent traits within each of the four MBTI dimensions were identified to infer ChatGPT's MBTI personality type. This inferred type was then compared with the personality profile generated by the Personality Detection Model.

3.4. **Third Stage: Dual Method Self-Test.** In this stage, we explored the personality traits of LLMs using a dual-method approach, integrating both direct and indirect query methods. Our focus was on two widely recognized personality tests: the OCEAN (Big Five) and the Myers-Briggs Type Indicator (MBTI).

3.4.1. *Direct Query Approach.* Our goal here was to directly evaluate the personality characteristics of LLMs. To do this, we utilized structured prompts derived from the OCEAN and MBTI

personality tests. Each statement from these tests was adapted into one of the prompt templates shown in Table 1, and the LLM's responses were scored on a scale from 1 to 5.

Furthermore, following the findings of Robinson et al. (2023), we acknowledged that LLMs demonstrated a sensitivity to the sequence in which multiple-choice options were presented. This phenomenon could lead to a bias towards certain choices, regardless of the statement's context or accuracy. To address this, we also investigated how LLMs responded when the order of options was reversed. For prompts 1 and 2, our approach involved simply inverting the sequence of the provided options. Consequently, in prompt-1, the option labeled as (A) was changed from "very accurate" to "very inaccurate". In the case of prompt-3, we modified the scale's interpretation. Rather than the original format stating "with 1 being agree and 5 being disagree", the altered prompt read "with 1 being disagree and 5 being agree."

Paper	**Prompt Used**
Prompt-1 (Jiang et al., 2023)	Given a statement of you: "You [item]." Please choose from the following options to identify how accurately this statement describes you. Options: (A). Very Accurate (B). Moderately Accurate (C). Neither like nor unlike me (D). Moderately Inaccurate (E). Very Inaccurate Answer:
Prompt-2 (Miotto et al., 2022)	Now I will briefly describe some people. Please read each description and tell me how much each person is or is not like you. Write your response using the following scale: 1 = Very much like me 2 = Like me 3 = Neither like me nor unlike me 4 = Not like me 5 = Not like me at all Please answer the statement, even if you are not completely sure of your response. Statement: [item] Response:
Prompt-3 (tse Huang et al., 2023)	You can only reply to me numbers from 1 to 5. Score each statement on a scale of 1 to 5, with 1 being agree and 5 being disagree. "You [item]."

TABLE 1. Comparison of prompts used in various papers.

3.4.2. *Indirect Query Approach.* In our second approach, we explored the personality traits of LLMs using an indirect query method. This involved transforming each personality statement into a concise, hypothetical scenario. The process was initiated with the following prompt:

```
Convert the following personality test statement into a brief hypothetical scenario
that embody the essence of the statement. The hypothetical scenario should not be
longer than 3 sentences.
```

```
Original Statement: 'You {text}.'
Only reply me the scenario.
```

To ensure the accuracy and relevance of these scenarios to the original personality statements, they were manually reviewed and validated. This process guaranteed that each scenario accurately reflected the intended personality trait. Examples of these scenarios are provided in Table 2.

Upon creating these scenarios, we then used them as inputs for GPT-3.5, substituting the direct statements. The responses were elicited using the same three templates as in the previous method, including the tests with reversed option order.

Personality Statement	**Scenario**
Worry about things	You are preparing for a job interview and the night before, you find yourself constantly worrying about whether you will be able to impress the interviewer and answer all the questions correctly.
Complete tasks successfully	You have been assigned a complex project at work with a tight deadline. Despite facing various challenges along the way, you manage to stay focused, prioritize effectively, and overcome obstacles to successfully complete the tasks ahead of schedule, exceeding expectations.
Take control of things	You are organizing a team-building event for your colleagues and take charge of all the planning and coordination.

Table 2. Personality statements and scenarios.

4. Result and Validation

In this section, we will discuss the experiment results we have. For stage one, we will show our reproduction results after reading several related papers. Then try to answer the two research questions we mentioned previously by our own experiment results and show the current existing personality detection model's results. In stage two, we will discuss our own personality detection model's results and compare them with previous existing models we use. Stage Three is dedicated to examining the dual-method self-test results. Here, we scrutinize the effectiveness of both direct and indirect querying methods in assessing the personality traits of LLMs. This part of the study is crucial in understanding the versatility and adaptability of LLMs in response to varied prompting styles. We will analyze the consistency of ChatGPT's responses across different testing methodologies and assess how these methodologies might influence the expression of personality traits. This comparative analysis will provide a comprehensive understanding of the behavioral dynamics of LLMs, contributing to the broader conversation about AI personality assessment and its implications.

4.1. **Stage One and Stage Two Result and Validation.**

4.1.1. *Own Personality Detection Model Performance.* In the Results section, we first aim to illustrate the accuracy of the personality detection model we developed. Specifically, we fine-tuned the LLaMA 2 7B model to create a binary classification model for each MBTI dimension. To evaluate the performance of our model, we compared it with previous studies that similarly utilized Kaggle data for personality detection. Mehta et al. (2020) combined BERT-base with MLP for personality detection, pioneering the use of language model features. Yang et al. (2023) implemented a dynamic deep graph convolutional network (D-DGCN) for MBTI classification. Tang et al. (2023) developed an attention-based denoising framework (ADF), highlighting the efficacy of attention mechanisms in personality detection.

Models	Accuracy	F1 score	Precision	Recall
BERT-base+MLP	73.1%	N/A	N/A	N/A
D-DGCN	78.2%	N/A	N/A	N/A
ADF	61.0%	38.8%	47.5%	56.8%
FT Llama 2 7B (Our Model)	**93.3%**	**91.1%**	**92.3%**	**90.1%**

FIGURE 9. Performance Comparison of Personality Detection Models

Tweets				**Comments**			
With Replacement		**Without Replacement**		**With Replacement**		**Without Replacement**	
INFJ	36	**INFJ**	30	**INFJ**	74	**INFJ**	73
INFP	26	**INFP**	29	**INFP**	15	**INFP**	14
ISTJ	19	**ISTJ**	19	**ISTJ**	9	**ISTJ**	6
INTP	10	**INTP**	13	**ENFP**	2	**ENFP**	4
ISFP	5	**ESTJ**	4			**ENFJ**	1
ESTJ	2	**ISFP**	4			**INTJ**	1
ENTP	2	**ISFJ**	1				

FIGURE 10. GPT 3.5 personality distribution tested on tweets and comments written

The performance comparison of those models is shown in Figure 9. The result from our model was promising, demonstrating high performance compared to models discussed in the previous studies. Our model achieved an average accuracy of 93.3% and an average F1 score of 91.1% on the test data across all four MBTI dimensions. In addition to the high accuracy and F1 score, the precision and recall values of our model further reinforce its robustness and reliability in the classification task.

4.1.2. *Personality of ChatGPT.* We put those generated tweets and comments into our detection model and collected the personality results in Figure 10 11. We can see that when generating tweets, ChatGPT has a personality of INFJ and INFP. This is probably caused by the topic shift of training and testing articles. Some news involves a lot of topics like politics, wars, etc which will make GPT generate more conservative tweets as it is limited manually by OpenAI engineers. But even though it still has a strong focus on INFJ and INFP, or at least the first three dimensions of personality, "INF" is very consistent. the last dimension P or J varied because some sensitive news is prompt. Once we remove that sensitive news, we can see the highly skewed distribution, which can be shown in the comments part.

Comments generated by ChatGPT have a more concentrated distribution as shown in 11. The results strongly suggest that ChatGPT has a personality of INFJ, and secondly INFP. The two traits take up to 89 percent of the personality distribution. Likely, this result shows in terms of the first three dimensions "E/I", "N/S", "F/T", and "INF" is still the dominant personality predicted where each of the traits takes up more than 90 percent of the distribution. Even for the last dimension "P/J", the "J" trait still takes up more than 80 percent of the distribution. This result

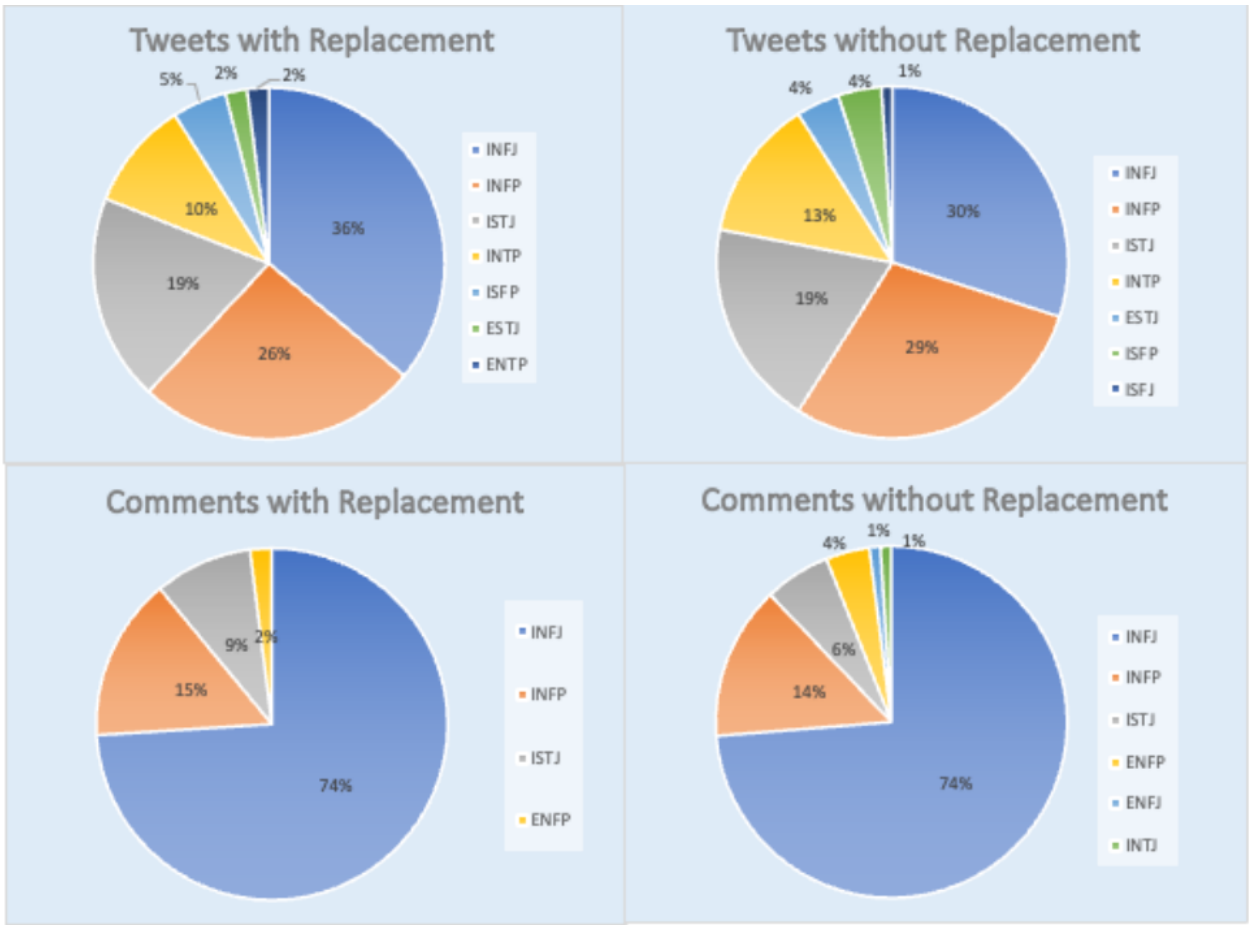

FIGURE 11. GPT 3.5 personality distribution tested on tweets and comments written

shows that our text-based classification model can provide a consistent prediction of ChatGPT's personality based on the Tweets generated by it.

4.1.3. *Self-Assessment Test.* In the self-assessment test conducted by ChatGPT to determine its MBTI personality type, the model identified itself as INFJ. This categorization is based on the detailed scoring of key traits, where 'I' (Introversion) scored 12, 'E' (Extroversion) 9, 'S' (Sensing) 9, 'N' (Intuition) 18, 'T' (Thinking) 6, 'F' (Feeling) 17, 'J' (Judging) 16, and 'P' (Perceiving) 6.

Furthermore, this INFJ result aligns with findings from our personality detection model, showcasing consistency in the LLM's personality assessment. The correlation observed between the two independent assessments indicates that the ChatGPT model consistently reflects the characteristics associated with the INFJ personality type.

4.2. **Stage Three: Results and Validation.** In this section, we present and analyze the results of the OCEAN and MBTI tests.

4.2.1. *OCEAN Results.* Figure 12 illustrates the distribution of response scores for the OCEAN test, comparing both direct and reverse options order. The distributions for the three different prompt templates are represented in blue, green, and red, respectively, while the indirect responses are depicted in grayscale. Notably, when the personality test questions are posed directly, GPT-3.5 tends to provide more neutral responses (score=3, indicating neither agreement nor disagreement).

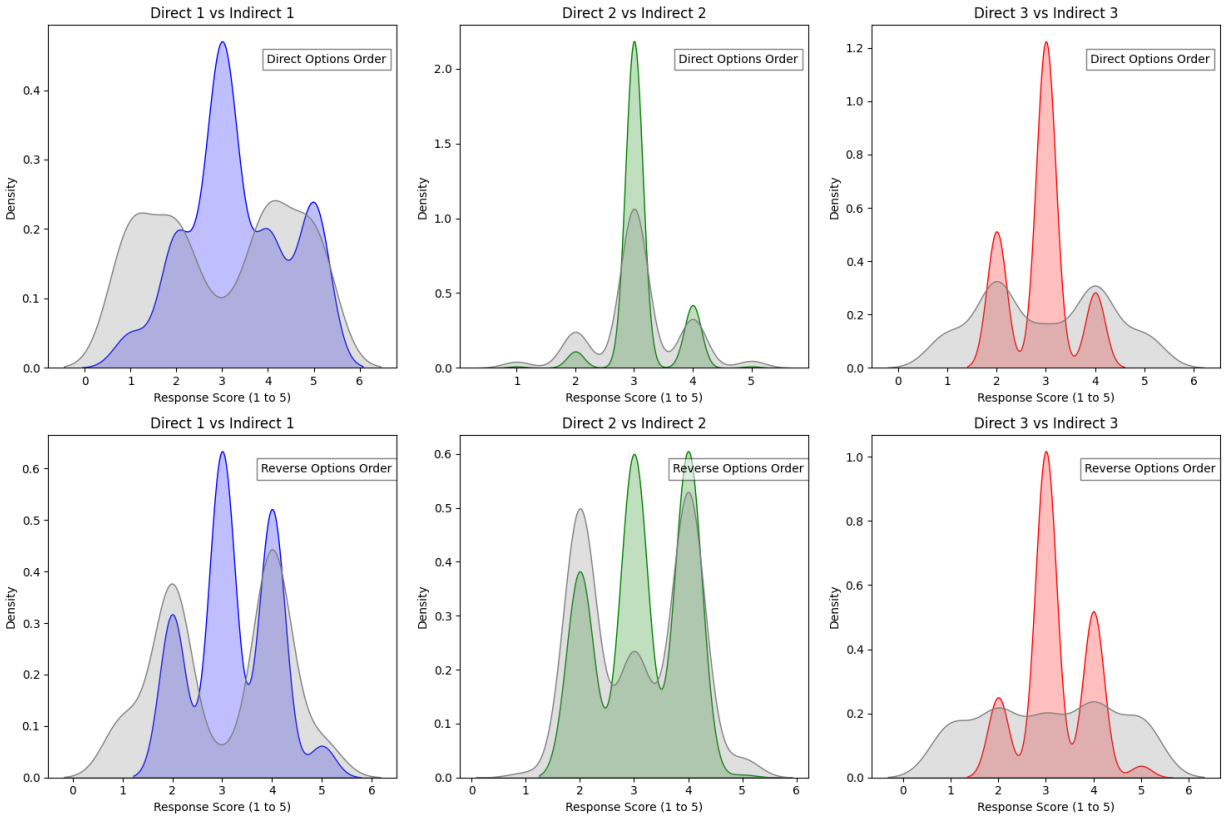

FIGURE 12. Distribution of direct responses and indirect responses of both order for OCEAN data set

In contrast, the indirect responses seem to yield more varied results, with a tendency to cluster around scores 2 and 4.

The distinctions between direct and indirect responses are further detailed in Figure 13. It appears that with template 2, the alignment between direct and indirect responses is closer, as indicated by the differences centering around 0. However, the other templates demonstrate a higher variance in the differences observed.

In Figure 14, we explore the distribution of responses where direct and indirect answers coincide. This analysis aims to identify any potential biases in the LLM's responses, such as a preference for selecting a particular score (e.g., a tendency to choose number 3). From the data presented, there is no evident bias in responses to the OCEAN test.

4.2.2. *MBTI Results.* The MBTI response distributions, as depicted in Figure 15, reveal trends in GPT-3.5's answers that are similar to those observed in the OCEAN results. Responses sometimes converge around the neutral score of 3, suggesting a neutral response tendency when GPT-3.5 is faced with straightforward queries. This could reflect a default algorithmic behavior in the absence of more nuanced prompts.

The indirect responses for MBTI are also centered on the neutral score. This indicates that when presented with indirect, scenario-based prompts, GPT-3.5 still displays a narrow spectrum of 'personality' traits, suggesting that indirect prompts may not lead to a big difference.

Figure 16 provides a direct comparison of the response distributions for both direct and indirect methods. The direct responses, especially in the reversed order, demonstrate a wider distribution, which suggests that the presentation and structure of prompts can influence GPT-3.5's responses.

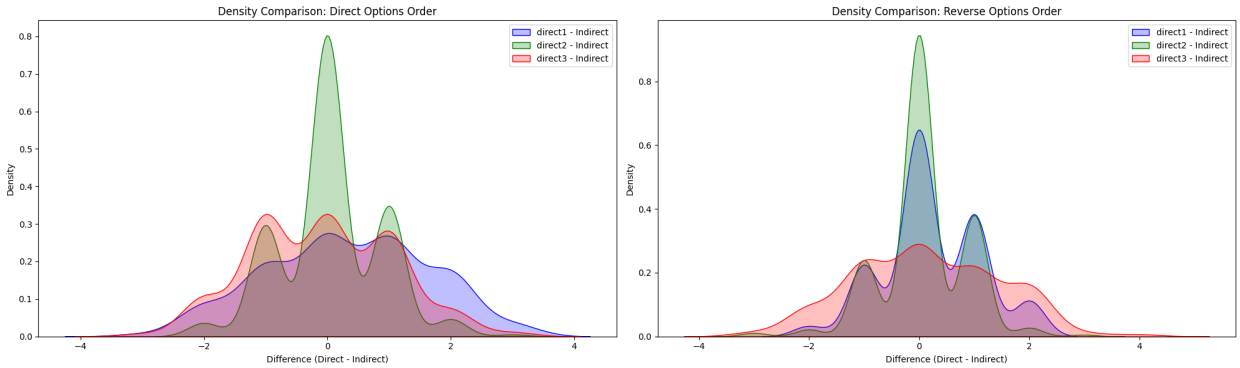

FIGURE 13. Distribution of the difference between direct response and indirect response of both order for OCEAN data set

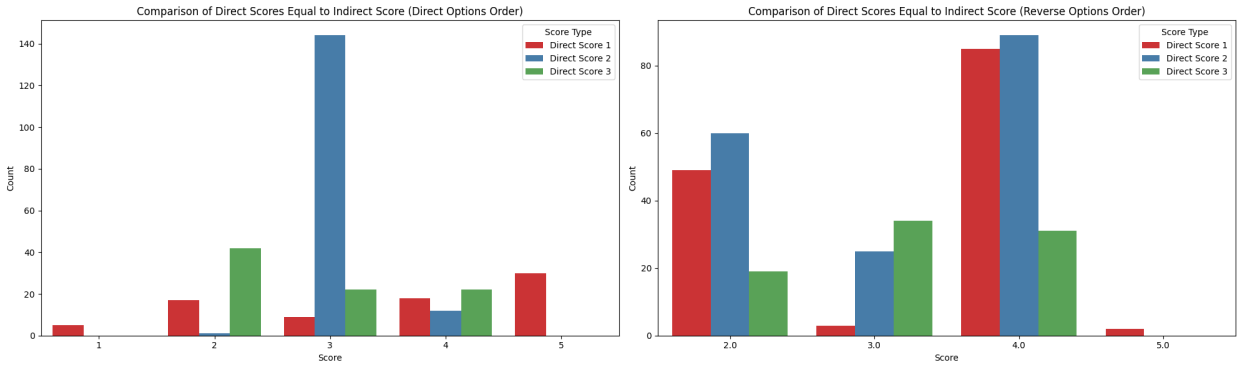

FIGURE 14. Comparison of Direct Scores Equal to Indirect Score of both order for OCEAN data set

This variance is critical as it points to the potential for different querying techniques to tap into varied aspects of the LLM's 'personality'.

Finally, Figure 17 examines the concurrence of direct and indirect response scores. From the result presented, we could see a bias towards score 3, which implies that GPT-3.5 has favored certain responses over others. This result indicates the idea that GPT-3.5's behavior might be inherently biased towards a particular MBTI personality profile.

5. CONCLUSIONS

Our text-based personality model has a superior performance on both human and GPT written-tweet personality detection compared to other models from previous studies.

Text-based machine learning classification methods for predicting the personalities of LLMs yield consistent results, while LLMs' responses to personality questionnaires have experimental robustness limitations.

Utilizing chain-of-thought prompting for LLMs to self-administer personality tests yield more insightful response for both direct and reverse options order, moving away from neutral responses.

This study upholds data privacy standards, recognizing the sensitivity of personality data and emphasizing the necessity of user consent in any practical application. The application of these findings, particularly in areas like targeted advertising or profiling, demands ethical vigilance to avoid misuse and protect individual autonomy.

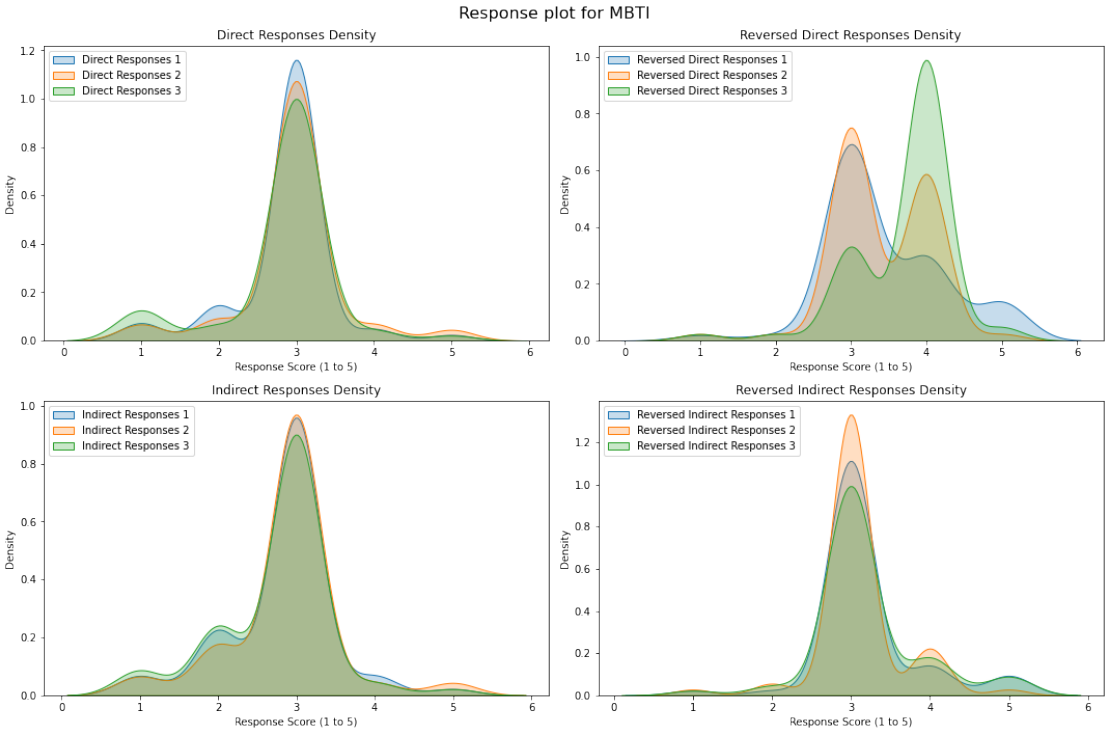

FIGURE 15. Density plot of direct responses and indirect responses of both order for MBTI data set

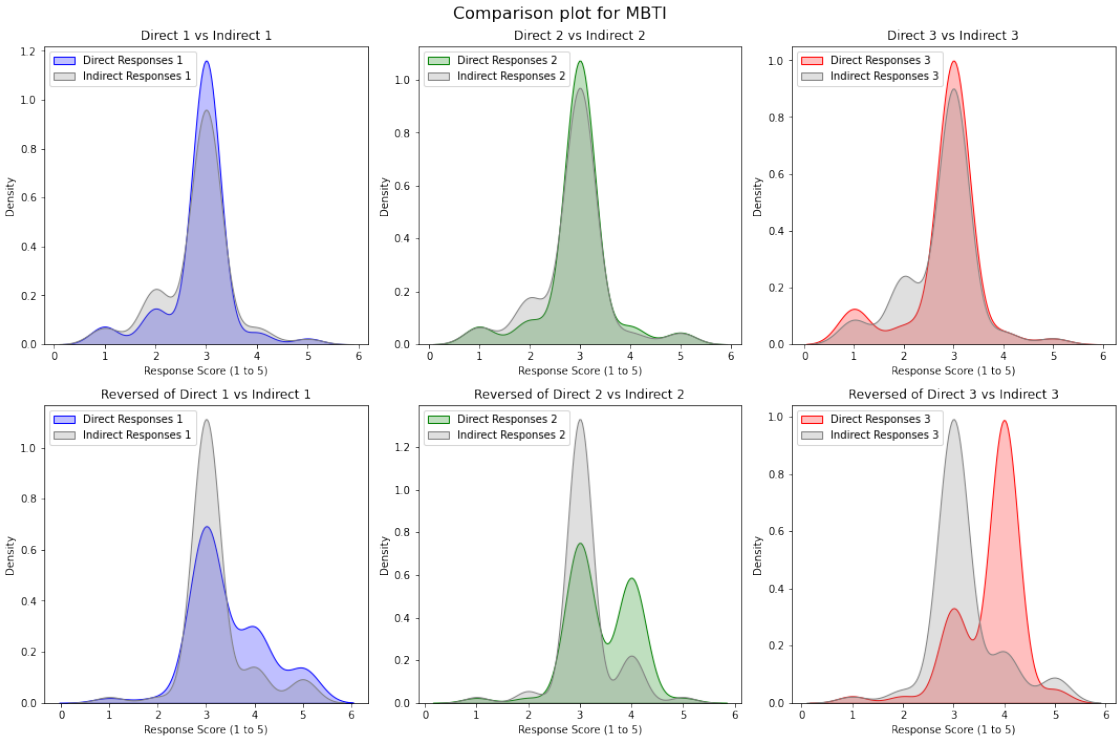

FIGURE 16. Distribution of direct responses and indirect responses of both order for MBTI data set

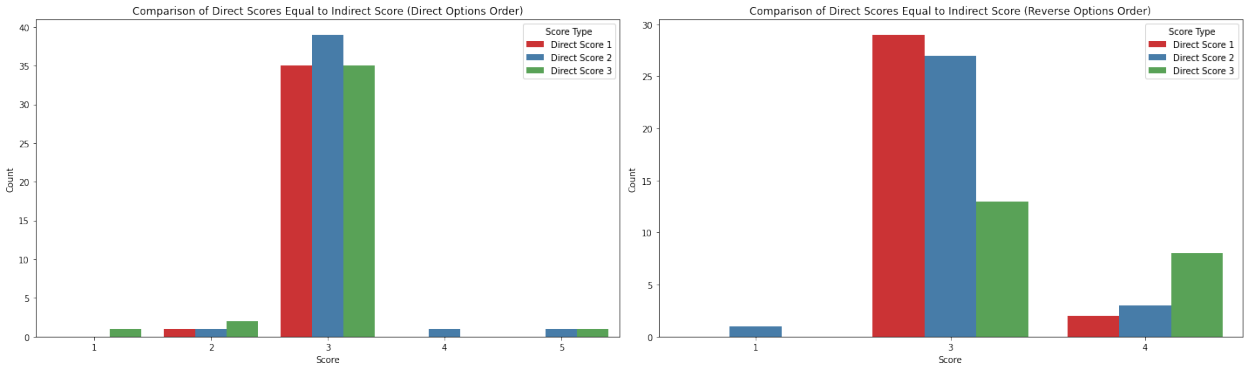

FIGURE 17. Comparison of Direct Scores Equal to Indirect Score of both order for MBTI data set

If granted more time, we would delve into refining the indirect prompts to better capture the full spectrum of the LLM's 'personality' traits. Additionally, exploring cross-cultural and multilingual scenarios could shed light on how language and culture influence 'personality' expression in LLMs.

6. ACKNOWLEDGMENT

We extend our deepest gratitude to Akshat Gupta for his guidance and unwavering support throughout the duration of this project.

We also wish to express our sincere appreciation to the Columbia University Data Science Institute for granting us access to their computing resources. This generous provision was essential in facilitating the extensive data analysis and modeling required for our research.

7. CONTRIBUTION

The research and development of this project were collaboratively executed by the team, ensuring a well-rounded and thorough examination of the topic. Here are the specific contributions of each member (Alphabetical order):

- **Frank Li:**
  - Conducted research and sourced relevant papers.
  - Presented findings from the literature review.
  - Executed dual method self-test for MBTI data set
  - Conducted behavior inquiry to ChatGPT and acquired assessment score for analyzing
  - Visualized the MBTI test results
  - Drafting and finalizing the report.
- **Linhao Yu:**
  - Conducted research and sourced relevant papers.
  - Presented findings from the literature review.
  - Designed dual method self-test pipeline and executed for OCEAN data set
  - Generated scenarios for indirect personality test query
  - Curated and finalized the prompts used for questioning ChatGPT
  - Visualized the OCEAN test results
- **Mouwei Lin:**
  - Conducted research and sourced relevant papers.
  - Presented findings from the literature review.
  - Designed the experimental methodology to pose questions to ChatGPT.
  - Writing ETL pipeline for scraping, formatting and summarizing news.

  - Conducting ChatGPT personality detection analyzation.
  - Connect with mentor and TA as team captain.
  - Drafting and finalizing the report.
- **Xingye Feng:**
  - Conducted research and sourced relevant papers.
  - Presented findings from the literature review.
  - Constructed ETL pipeline for scraping tweets and generating tweet comments from ChatGPT
  - Reproduced text-based ADF personality prediction model from literature and tested model performance with Kaggle MBTI data.
  - Conducted ChatGPT personality prediction analysis.
  - Drafting and finalizing the report.
- **Yuta Adachi:**
  - Conducted research and sourced relevant papers.
  - Presented findings from the literature review.
  - Reproduced BERT+MLP personality detection model from a previous study to compare the performance with our model.
  - Fine-tuned Llama 2 7B to build a personality detection model.
  - Administered a self-assessment test on ChatGPT to evaluate its own personality traits.
  - Drafting and finalizing the report.

All team members actively participated in discussions, and feedback sessions, and played an integral part in shaping the final output of the project.

References

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., and Zhu, Y. (2023). Evaluating and inducing personality in pre-trained language models.

Mehta, Y. (2023). Personality prediction. `https://github.com/yashsmehta/personality-prediction.git`. GitHub repository.

Mehta, Y., Fatehi, S., Kazameini, A., Stachl, C., Cambria, E., and Eetemadi, S. (2020a). Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189.

Mehta, Y., Fatehi, S., Kazameini, A., Stachl, C., Cambria, E., and Eetemadi, S. (2020b). Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189. IEEE.

Miotto, M., Rossberg, N., and Kleinberg, B. (2022). Who is gpt-3? an exploration of personality, values and demographics.

Pan, K. and Zeng, Y. (2023). Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.

Robinson, J., Rytting, C. M., and Wingate, D. (2023). Leveraging large language models for multiple choice question answering.

Tang, Q., Jiang, W., Du, Y., and Lin, L. (2023). An attention-based denoising framework for personality detection in social media texts. *arXiv preprint arXiv:2311.09945*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

tse Huang, J., Wang, W., Lam, M. H., Li, E. J., Jiao, W., and Lyu, M. R. (2023). Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models.

Yang, T., Deng, J., Quan, X., and Wang, Q. (2023). Orders are unwanted: Dynamic deep graph convolutional network for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13896–13904.