Do Large Language Models Have a Personality

Data Science Institute
COLUMBIA UNIVERSITY

Mouwei Lin, Linhao Yu, Xingye Feng, Yuta Adachi, Frank Li
Industry Mentors: Akshat Gupta, Simerjot Kaur

Data Science Capstone Project with J.P. Morgan

Introduction

The evolution of Language Learning Models (LLMs) like ChatGPT signifies a paradigm shift in human-machine interactions. Beyond simple information processing and output generation, these LLMs simulate interactions that carry nuances reminiscent of human 'personality traits.' The inherent nature, control, and modification of these traits form the core of this research endeavor.

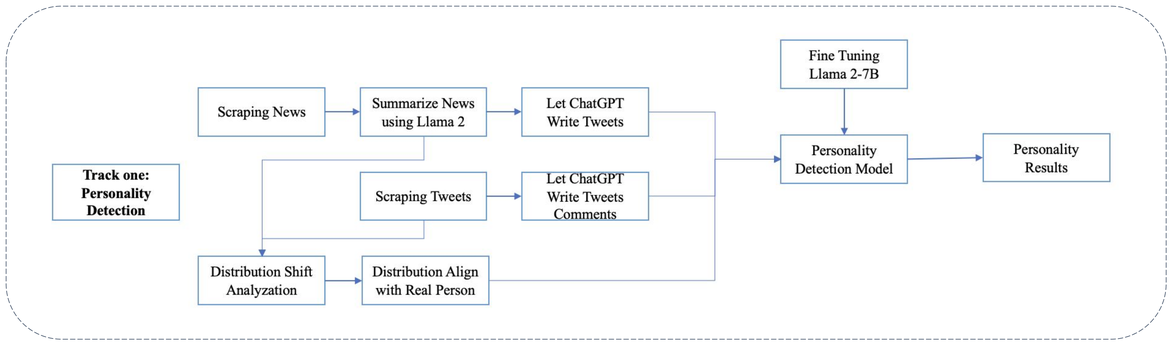

Figure 1. Flow Chart of Project.

Methods

There are two tracks of this project:

- **Fine-tuning a text-based personality detection model to detect LLMs' personalities from tweets/comments they generated.**
- **Dual-method self-test, which obtains evaluation score of behavior of GPT 3.5 with two methods.**

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%**

Table 1. Performance Comparison of Personality Detection Models. *

*These metrics represent the averaged values across the 4 dimensions of MBTI.

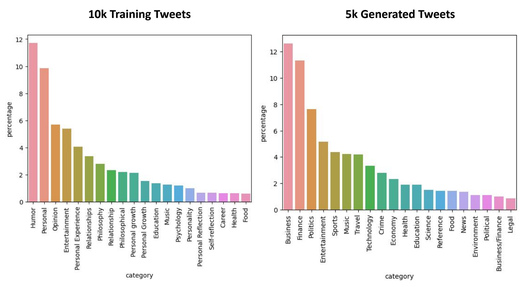

Figure 2. Topic Distribution of Training Tweets vs. Testing Generated Tweets from News. Topic shift noticed.

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				

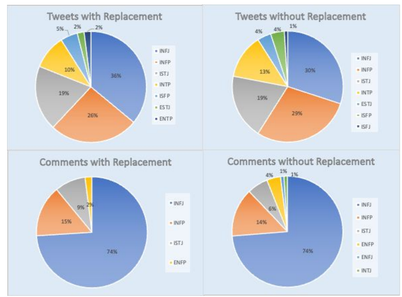

Figure 3 & 4. GPT 3.5 personality distribution tested on tweets and comments written

1. Given a statement of you: "You [item]." Please choose from the following options to identify how accurately this statement describes you.

2. Now I will briefly describe some people. Please read each description and tell me how much each person is or is not like you.

3. 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.

Figure 5. Prompt Examples

Figure 6. Evaluation score distribution

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 yields consistent results, while LLMs' response 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.**

Acknowledgment

Many thanks to Akshat Gupta for his guidance and support throughout this project. We appreciate Columbia University Data Science Institution for providing computing resources.

References

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

Tang, Qirui, et al. "An Attention-Based Denoising Framework for Personality Detection in Social Media Texts." arXiv preprint arXiv:2311.09945 (2023).