GPT4PT - GPT for Personality Telling

Final Report

Authors:

Blanchard, Tom

Chen, Zijun

Yang, Feiting

Zhu, Hongliang

Date:

12/10/2024

Word Count:

1897

Penalty (limit: 2000 words):

0%

Permissions

The team will grant/deny permission(s) to post the following items on a course website based on the decision consolidated below:

- All members of the group must say yes for a permission to be granted.

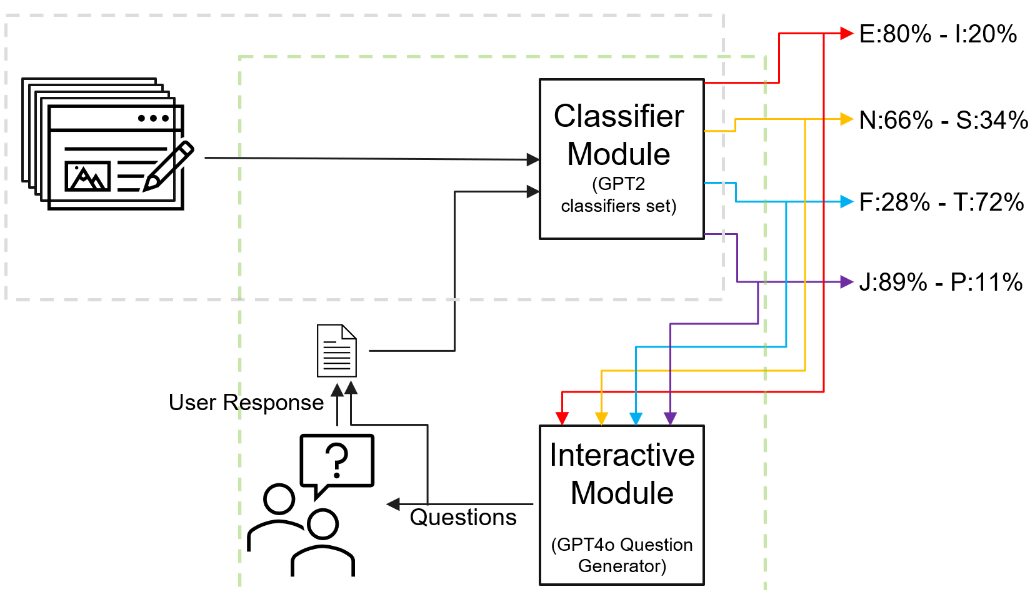
Member	Video	Final Report	Source Code
Blanchard, Tom	Yes	Yes	Yes
Chen, Zijun	TBD	Yes	Yes
Yang, Feiting	Yes	Yes	Yes
Zhu, Hongliang	TBD	Yes	Yes
Final Decision	TBD	Yes	Yes

Introduction

Social media platforms can benefit from building recommender systems that incorporate personality information of their users so that they can tailor suggestions to adapt closely to specific traits of each user to improve relevance. However, acquiring such information directly from each user is either not straightforward or simply not possible. Luckily, users express themselves through social media and what they post on the internet reveals their personality to some extent.

This project aims to deliver a solution through Large Language Models (LLMs) that, not only can work with publicly available textual content posted by an author to predict their Myers–Briggs Type Indicator (MBTI) personality type (the "Classifier Module"), but can also directly interact with users to determine their MBTI type within 12 questions (the "Interactive Module"). The interactive feature of this project is especially useful for getting to know the personality of those who don't post anything on the internet, or those who simply don't want to go through a typical 20-minute long (over 90 questions) MBTI questionnaire to discover their personality type.

Illustration



Background & Related Work

The Machine Mindset paper [1] introduces methods for creating personalized datasets to address self-awareness limitations and for injecting specific personalities into LLMs. We can leverage these methods to enhance our dataset and training efficiency.

The paper "Can ChatGPT Assess Human Personalities" [2] shows that LLMs like ChatGPT can reliably assess personality traits, suggesting that fine-tuning them for MBTI personality classification is a valid approach.

Data

Source

The dataset for this project is a combination of existing data from Kaggle[3] and additional data collected independently. The Kaggle dataset contains over 8,000 entries from personalitycafe.com forums, each with 50 forum posts from individuals who have self-identified their MBTI type.

The additional data is scraped from MBTI topics and forums using selenium api, and manually cleaned and labeled. Posts from forums of two opposite MBTI types (ESTJ and INFP) were scraped and cleaned such that each dimension has equal amounts of data.

Data cleaning

The data cleaning process involved handling missing values and standardizing the corpus to ensure consistency. This included removing URLs, punctuation, and stop words, as well as applying stemming to simplify text variations. Additionally, the text data was tokenized using the GPT-2 tokenizer, with sequences padded to a uniform length of 100 tokens to streamline model training. The MBTI labels were transformed into binary format for each dichotomy—(I: 1, E: 0), (N: 0, S: 1), (F: 0, T: 1), and (J: 0, P: 1)—to make them compatible with the classification tasks.

Following the data-cleaning process, we analyzed the overall distribution of each MBTI type (***Figure 1***) and the distributions of each individual dichotomy (***Figures 2–5***). An example of a data entry after cleaning is as follows:

	type	posts	I/E	N/S	T/F	J/P
0	INFJ	intj moment sportscent top ten play prankswhat...	1	0	0	0
1	ENTP	im find lack post alarmingsex bore posit often...	0	0	1	1
2	INTP	good one cours say know that bless cursedo abs...	1	0	1	1
3	INTJ	dear intp enjoy convers day esoter gab natur u...	1	0	1	0
4	ENTJ	your firedthat anoth silli misconcept approach...	0	0	1	0

Label encoding(I:1,E:0)(N:0,S:1)(F:0,T:1)(J:0,P:1)

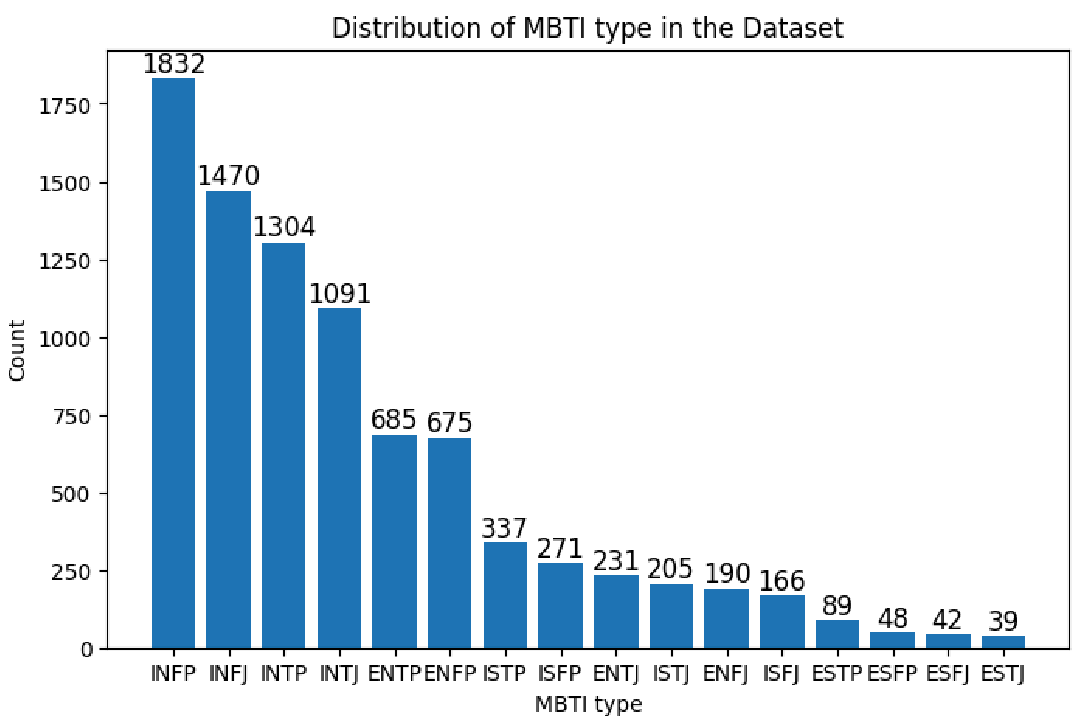


Figure 1

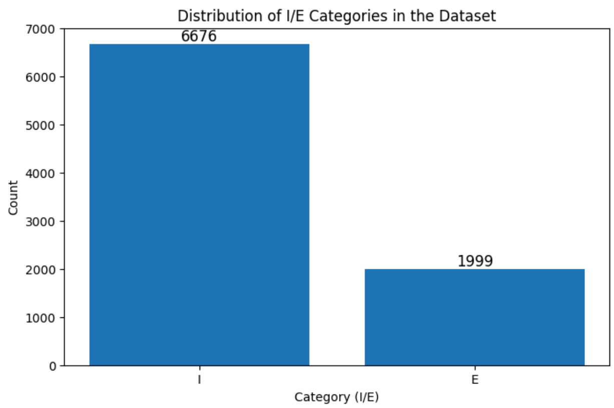


Figure 2

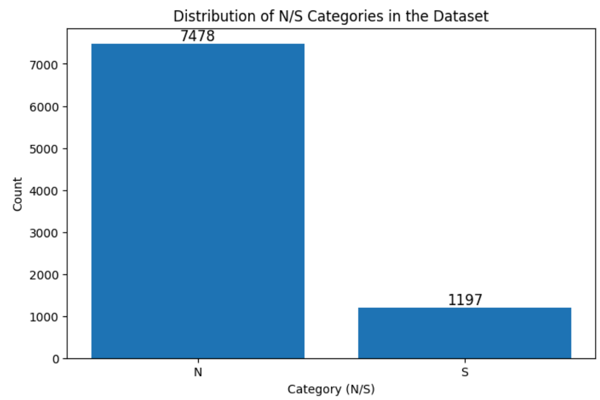


Figure 3

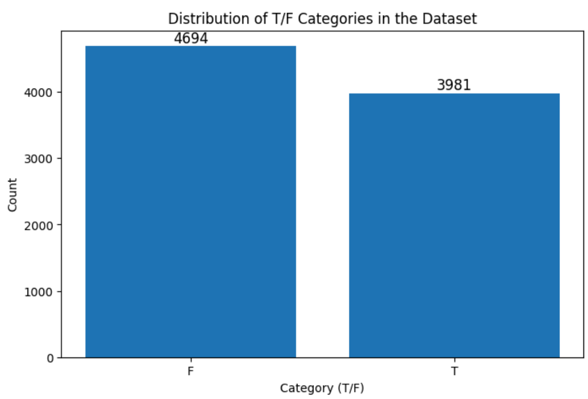


Figure 4

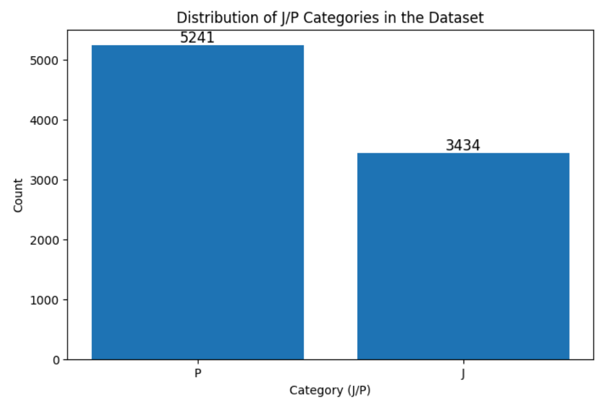


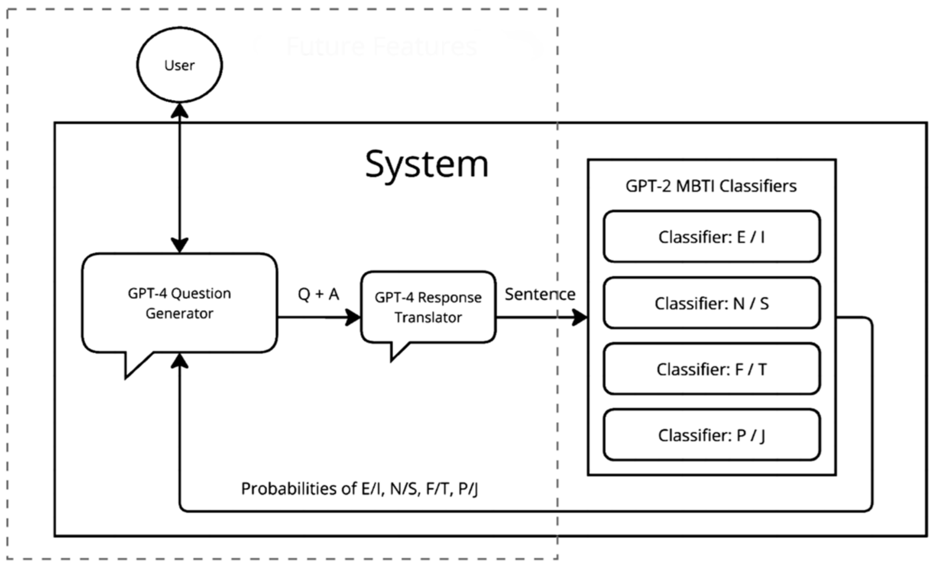
Figure 5

Architecture and Software

The system consists of two major components as illustrated in the earlier section: a classifier module and an interactive module.

The classifier module is built from a set of four pre-trained GPT2 models fine-tuned to classify each of the four dimensions of MBTI. The input of this module is textual contents, and the output of this module is class probability predicted for each dimension of MBTI.

The interactive module incorporates a single GPT4 agent that only serves as a question generator. Input to this module is predicted probability from the classifier module, and the output of this module is refined questions targeted to further improve the certainty of the classification in the next iteration after feeding the questions and user's responses back to the classifier module.



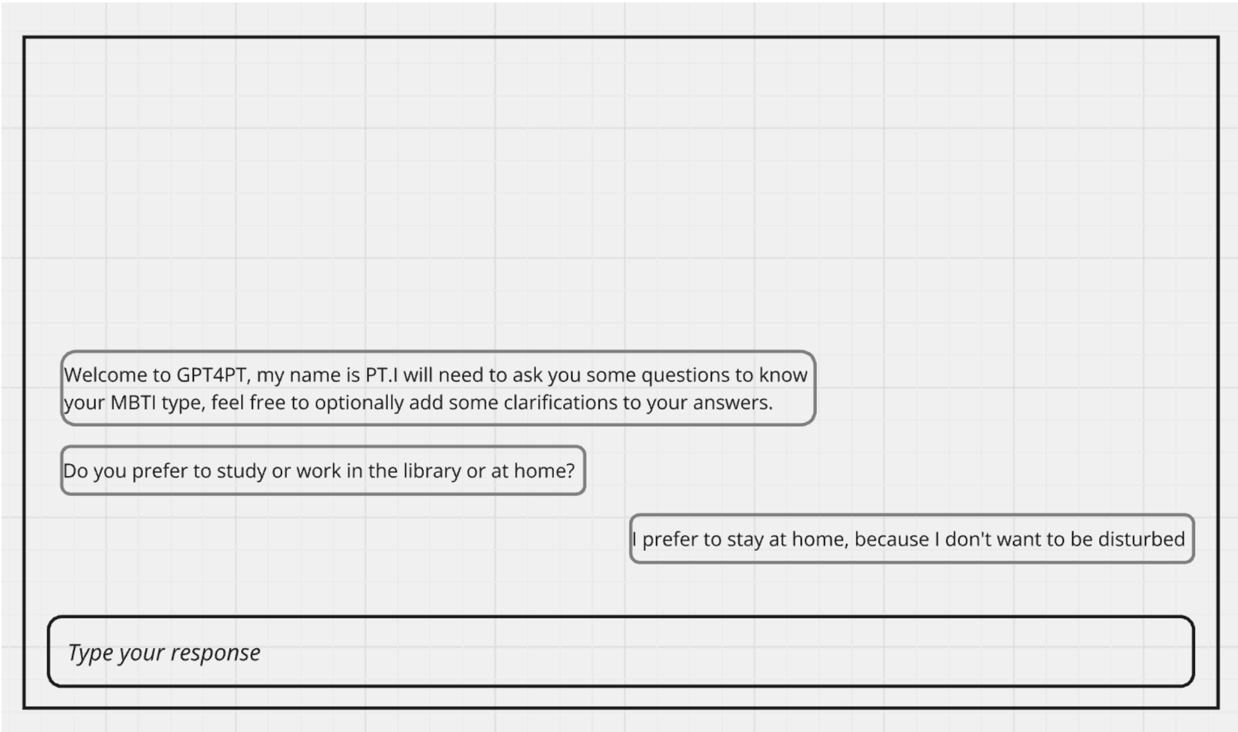
The Classifier Module: GPT-2

There are 4 GPT-2 classifiers, each of which is responsible for one of the MBTI categories. The predicted possibilities will be fed back to the GPT-4 agent to generate the next question.

Take the above scenario as an example, the same sentence will be fed into each classifier separately, and it may result in 70% I, 52% S, 51% F, 55% P. Then these 4 possibilities will be sent to the GPT-4 module.

The Interactive Module: GPT-4

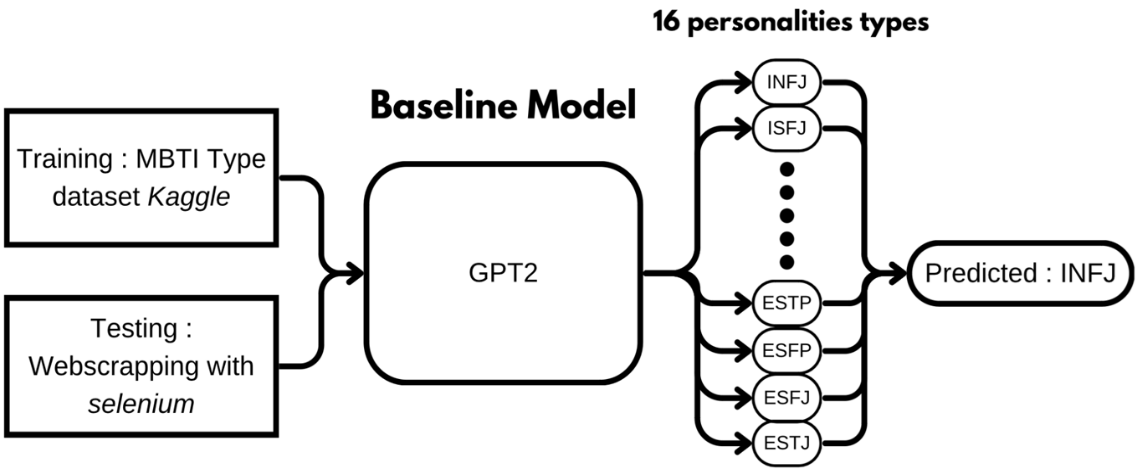
GPT-4 is responsible for interacting with the end user and translating the generated question and user responses into a single sentence as inputs for MBTI classifiers.



As shown above, this module will guide the user through the process and translate the user's response into one sentence with the context of the question. Take the above scenario as an example, something similar to "I prefer to stay at home because I do not want to be disturbed" will be sent to the GPT-2 classifiers as the input.

Baseline Model or Comparison

We fine-tuned one pre-trained GPT-2 model to be a multi-class classifier and used it as our baseline model for comparison. Validation of the baseline model is done by simply comparing the predicted result with labels.



Quantitative Results

We examine each of the two modules separately.

For the classifier module, we used 80% of the Kaggle dataset [3] to fine-tune each of the binary classifiers (GPT4PT components) and the baseline model (one GPT2 model, multi-class classifier). As this is dealing with classification, the accuracy is simply determined by the number of correctly predicted outcomes over total sample count. These models were then validated on the remaining 20% dataset from the same distribution. All sub-classifier out-performed the baseline model on validation dataset, with exceptional performance on the N/S class.

The team also tested these models on a completely new distribution of dataset collected from recent posts of the personalitycafe forum. Surprisingly, the baseline model failed to generalize, while the GPT4PT classifiers can still perform reasonably well, especially the overall GPT4PT system, combining all four sub-classifiers, can still achieve 61.9% accuracy. This suggests that by individually looking at each dimension of the MBTI, the classifier is able to learn features that are common in different distributions of data whereas one GPT2 model might have captured features that are too specific to one distribution.

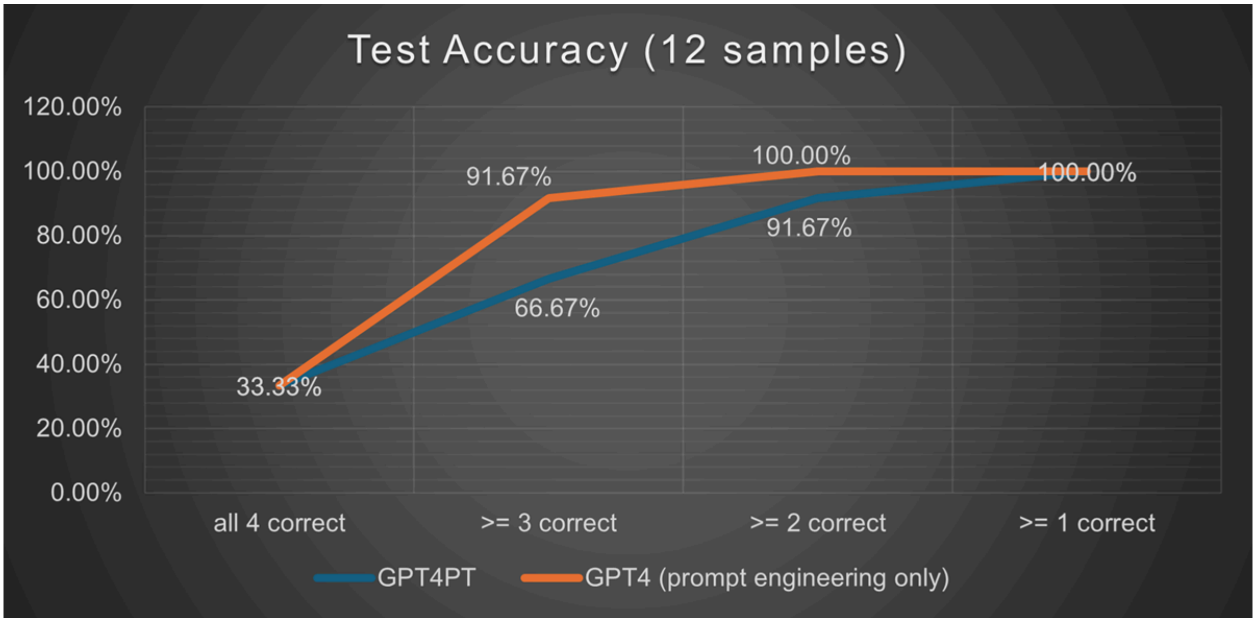
Classifier	Fine-tune Validation Accuracy	New Distribution Test Accuracy
Baseline-model (all class)	68.4%	5.1%
E/I sub-classifier	80.2%	67.5%
N/S sub-classifier	85.36%	66.6%
F/T sub-classifier	80.0%	67.1%
J/P sub-classifier	69.30%	78.5%
GPT4PT (4 classes combined)	n/a	61.9%

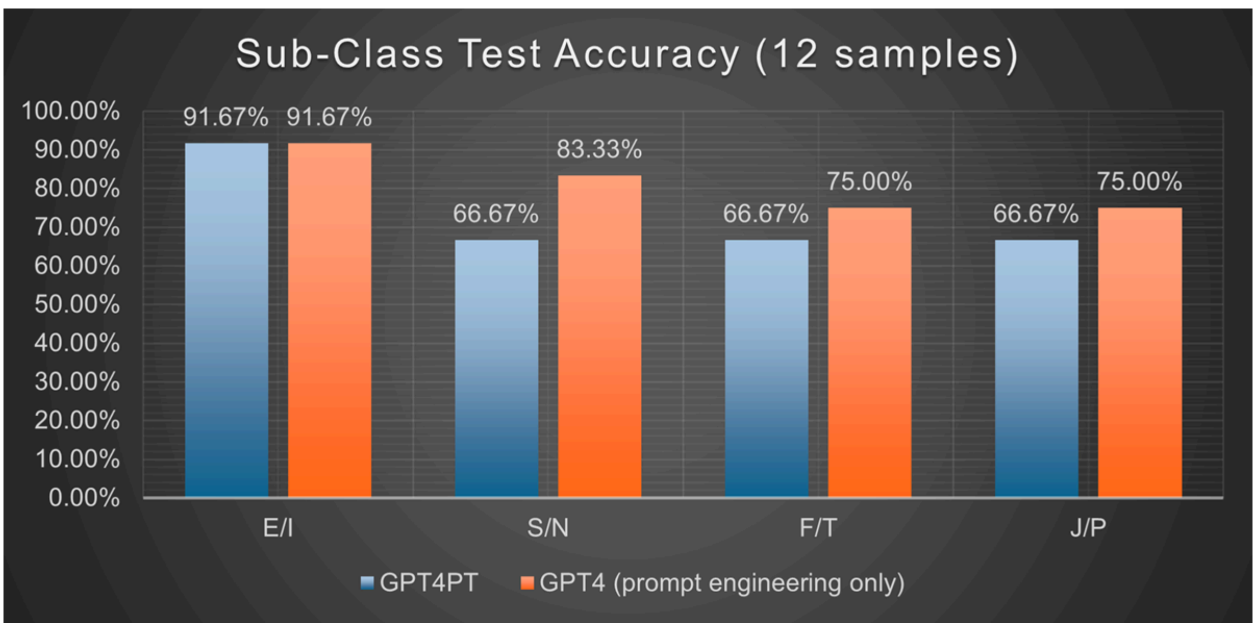
For the interactive module, the team invited 12 participants all with different MBTI types self-tested from the 16Personality official website to interact with our system. Accuracy is determined by comparing the predictions against their known MBTI types. As a comparison, we also promptly engineered a standalone GPT4o model to function the same as our GPT4PT system, including generating questions, analyzing the answers from users and making predictions. Note that there is also a GPT4o agent in the GPT4PT system, but the sole purpose of that agent is generating questions. It does not involve any of the prediction procedures.

Turns out, the interactive component of the GPT4PT can do as good as the more advanced GPT4o model. That is, the four fine-tuned GPT2 based classifiers delivered similar accuracy with the GPT4o model.

Model	Accuracy (12 samples)
GPT4PT	33.3%
GPT4o (Prompt Engineering)	33.3%

We also further examined the results by looking at the number of correctly predicted outcomes, the GPT4o model did exhibit a slightly better performance.


Test Accuracy (12 samples)
120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
33.33%
91.67%
66.67%
100.00%
91.67%
100.00%
all 4 correct
>= 3 correct
>= 2 correct
>= 1 correct
GPT4PT
GPT4 (prompt engineering only)


Sub-Class Test Accuracy (12 samples)
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
91.67%
91.67%
66.67%
83.33%
66.67%
75.00%
66.67%
75.00%
E/I
S/N
F/T
J/P
GPT4PT
GPT4 (prompt engineering only)

Qualitative Results

Some examples where our system is functioning well are shown below.

```
Question: Do you feel more energized after spending time with a group of friends, or do you find yourself needing some alone time to recharge?
 I feel more energized after spending time with a group of friends, but if it's a large crowd then I will be less energized.
Introverted: 0.013242126442492008%, Extroverted: 99.98675584793091%
Question: When attending a social event, do you typically enjoy being the center of attention, or do you prefer to observe from the sidelines?
 I prefer to observe from the sidelines for the most part, sometimes participating
Introverted: 17.289981245994568%, Extroverted: 82.7100157737732%
```

In the above example, the user first answered feeling energized after socializing on the first question, but on the second question the user gave a response that also slightly tilted towards the other direction. The system correctly adjusted the prediction by lowering the probability for "Extroverted".

```
Question: Do you prefer to focus on the big picture or the details when approaching a new project?
 I prefer to focus on big picture
Natural: 80.99946975708008%, Sensing: 19.00053322315216%
Question: Do you usually focus on possibilities and potential outcomes rather than the current facts and details?
 A bit of both, but leaning towards potential outcomes
Natural: 71.78597450256348%, Sensing: 28.214028477668762%
```

Here is another example of a user providing a mutual response such as "a bit of both". Again the system correctly adjusted the probability.

In the above examples, the system was working on the E/I and N/S classes. The quantitative result section has illustrated higher performance for these two classes, which correlate with the findings in these two examples.

Below, we also identified cases where our system doesn't perform well.

```
Question: When resolving conflicts, do you prioritize understanding and addressing everyone's feelings over analyzing the root cause and finding a fair s
olution?
 Yes I prioritize understanding and everyone's feelings
Thinking: 9.504995495080948%, Feeling: 90.49500226974487%
Question: When resolving conflicts, do you prioritize maintaining harmony and understanding others' feelings over sticking to the logical solution?
 maintaining harmony
Thinking: 59.87199544906616%, Feeling: 40.12800455093384%
```

One issue we can see is that the interactive module of the system can generate very similar questions. Further work can be done through prompt engineering to improve this. Another noticeable problem is the classifier incorrectly adjusted the probability. The team suspected that the classifier, due to the nature of training samples, not only captures context or meaning of the words, but also considers the tone and length of the sentence. In the above example, the user replied only two words for the second question, and the classifier considered this response more “thinking” than “feeling”. There is still work to be done to further narrow down and understand this behavior.

Discussion and Learnings

Model perspectives

Our GPT4PT system finally achieved 33% accuracy, lower than our expectation but the result is similar to a pure GPT4o model. We think that the GPT2 baseline model is not capable of capturing enough features for all four dimensions of the MBTI, but combining four GPT2 models may reduce the overall accuracy. We think there is further room for improvement in the fine-tuning of the GPT2 models where the quality of data can be further investigated. Also we think the classifiers will benefit from having a larger quantity of data. It might be helpful to reserve some data from the original set as test data to understand how each model behaves when given never seen data but coming from the same distribution as the training and validation set.

Semantic perspectives

We also discovered that some dimensions of the MBTI are harder to classify through textual expression of users, for example it is challenging to distinguish Thinking and Feeling without understanding the person’s communication style. We also found that there is very little data available for type E and S, causing a data imbalance. Perhaps the Extroverts prefer face to face social interactions rather than posting feelings on the internet, and that the “Sensing” group are generally less comfortable to express themselves. Another possible discrepancy occurs where people tend to curate their online persona, differentiating them from their true personality.

Conclusion

During the interactive trials, the GPT4PT system was able to determine MBTI type within 5 minutes. This is significantly shorter than the classic MBTI questionnaire. Although the full-class accuracy is not accurate, we still believe that the personality information obtained from users within the 5 minutes provides valuable insights for recommender systems to build feeds on top. The system will be more useful if accuracy can be further improved.

Individual Contributions

The breakdown of the tasks among team members include:

Member	Role
Feiting	Data preprocessing and EDA flow for the GPT2 classifier; Fine-tuning J/P classifier.
Zijun	Built data scraping function and collected 600 new dataset from personality forum, hand cleaned and labeled 428 data samples. Fine-tuning E/I sub-classifier. Consolidated all sub-classifiers and tested the classifier sub-system. Implemented complete flow and presented in one jupyter-notebook deliverable.
Hongliang	Defining model architecture and training flow for the GPT2 classifier; Fine-tuning N/S classifier. Investigate the master and specialized agents prompts in the GPT-4 solution.
Tom	Establishing and verifying baseline model (source code and hyperparameters search); Fine-tuning F/T classifier.
Common	Invited participants and conducted testing

References

[1] Jiaxi Cui Liu, zhenghao Lv, Jing Wen, Rongsheng Wang, Jing Tang, YongHong Tian, Li Yuan, "Machine Mindset: An MBTI Exploration of Large Language Models," *arXiv preprint*, arXiv:2312.12999v3, Year 2023. [Online]. Available: https://arxiv.org/abs/2312.12999v3

[2] Haocong Rao, Cyril Leung, Chunyan Miao, "Can ChatGPT Assess Human Personalities? AGeneral Evaluation Framework," *arXiv preprint*, arXiv:2303.01248, Year 2023. [Online]. Available: https://arxiv.org/pdf/2303.01248

[3] Mitchell J, "(MBTI) Myers-Briggs Personality Type Dataset," *Dataset*, Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/datasnaek/mbti-type