

# Personality AI

2019170609 정진성

2018320203 오윤석

2022320076 신동인

## Abstract

인간의 성격은 개인의 행동, 사고, 감정 등을 묘사하는 기본적인 특성이다. 최근 몇 년 간, 연구자들은 성격을 가치 있는 요소로 생각하여 다양한 작업(감정 분석, 정신 건강 분석 등)에 통합하였고, 이를 통해 상당한 성능 향상을 이루었다.

대규모 언어 모델 (Large Language Models, LLMs)의 발전에 따라, 텍스트 기반 성격 인식을 위해 LLMs 기반의 방법들이 제안되고, 높은 성능 향상을 이루었다. 또한 LLMs는, 단순한 질문에 적합한 대답을 하는 것을 넘어서 인간의 대화내용, 긴 맥락이 있는 대화 속에서 상대의 성격을 파악하고, 이에 따라 적절한 응답을 할 수 있으며, 사용자가 요구하는 성격 유형에 따라 적절한 응답을 바꿀 수 있는 수준까지 발전했다.

이번 연구에서는 LLMs이 충분한 데이터를 학습하여, 인간의 성격을 예측하는 능력을 가지고 있고, 적절한 방법을 통해 LLMs에게 성격을 주입할 수 있음을 확인한다. 또한, 추가적인 데이터의 학습 없이 프롬프트 기법을 바꾸는 프롬프트 엔지니어링 만으로 성격 예측과 주입의 성능을 향상할 수 있음을 확인한다. 이를 바탕으로 인간과 AI의 상호작용간 성격 예측과 주입에 따른 반응을 통해 사용자의 신뢰를 높일 수 있음을 시사한다.

## 1. Introduction

대규모 언어 모델(LLMs)의 급속한 발전은 자연어 처리 분야에 혁신을 가져왔으며, 특히 Few-shot 및 Zero-shot 학습 능력을 가능하게 했다 [1]. 이러한 Few-shot 및 Zero-shot 학습 능력은 LLMs가 제한된 데이터나 명시적인 지침 없이도 다양한 작업을 수행할 수 있게 해주었다. 이는 LLMs가 인간과 유사한 방식으로 언어를 이해하고 생성할 수 있는 잠재력을 가지고 있음을 시사한다. 이러한 발전은 인간과 AI 사이의 상호작용에 중요한 영향을

미치고 있으며, 특히 LLMs의 성격 예측 및 주입 능력은 인간과 AI 사이의 신뢰를 높이는 데 상당한 잠재력을 가지고 있다.

최근 연구에 따르면 방대한 양의 사용자 행동 및 언어 데이터를 학습한 LLMs [1]는 인간과 유사한 행동을 보일 수 있으며 [2], 심지어 인간의 "마음"을 이해하는 데 있어 인간을 능가할 수 있다 [3]. 이는 LLMs가 사용자의 성격을 정확하게 예측하고 이해할 수 있는 능력을 시사한다. 실제로 최근 연구는 LLMs가 성격 특성을 효과적으로 통합하고 [4], [5] 성격을 정확하게 예측할 수 있음을 입증했다 [6].

LLMs에 성격 예측 및 주입 기능을 통합하면 인간과 인공지능 사이의 신뢰 간극을 좁힐 수 있는 잠재력이 있다. 사용자 성격을 예측하고 이해하는 능력은 LLMs가 개인의 선호도에 더 잘 맞도록 의사소통 스타일, 어조 및 내용을 조정할 수 있게 한다[6]. 또한, 사용자가 원하는 성격 특성을 LLMs에 주입할 수 있게 함으로써, AI는 사용자의 기대에 맞춰 행동과 반응을 조정할 수 있어 더욱 조화롭고 만족스러운 상호작용을 만들어낼 수 있다[4], [5].

이러한 개인화는 인간-AI 상호작용에서 친밀감과 편안함을 증진시켜, 사용자가 AI 시스템을 더 신뢰하고 참여하도록 만들 수 있다. 인간의 성격 특성을 이해하고 모방할 수 있는 AI 시스템을 만들어 더 자연스럽게 직관적인 의사소통 채널을 구축할 수 있으며, 이는 사용자가 AI를 더 친근하고 공감할 수 있는 존재로 인식하게 되어 신뢰도를 높일 수 있다.

본 연구는 LLMs의 성격 주입과 성격 예측을 통해 인간과 AI 사이의 신뢰를 높이는 방법에 대해서 연구하기에 앞서, LLMs의 성격 주입과 성격 예측의 성능을 확인하여 향후 연구의 베이스라인을 제시하기 위해 다음과 같은 실험을 진행한다.

첫번째로, LLMs이 충분한 데이터를 학습하여, 인간의 성격을 예측하는 능력을 가지고 있음을 확인하고, 프롬프트 기법을 바꾸는 것 만으로 성격 예측 성능을 향상할 수 있음을 확인한다.

두번째로, 프롬프트 엔지니어링을 통해 효과적으로 LLM에게 성격을 주입하는 방법론을 바탕으로, 이를 구현하여 각 모델과 방법론에 따른 성격 주입결과를 분석한다. 이를 통해, 특정 성격을 주입하고 유도할 수 있음을 확인하여, 인간과 AI 간의 상호작용에서 성능을 크게 향상시킬 수 있고, 다양한 분야에서 활용될 수 있음을 확인한다.

이러한 연구를 바탕으로 추후 LLM과 인간의 상호작용에서 인간의 성격을 예측하고 그에 따른 성격 주입과 행동 유도를 통해 인간과 AI 간의 신뢰를 높일 수 있음을 증명한다.

## 2. Personality Injection

### 2-1. Related Work

LLM에게 성격을 주입하고 이 결과를 분석하기 위해서는, 인간 행동을 정량화하고 분석하는 것과 마찬가지로, 성격 유형을 측정하고 정량화 할 수 있는 표준화된 방식이 필요할 것이다. Guangyuan Jiang의 연구(2023)에서는 Machine Personality Inventory(MPI)와 Personality Prompting(P2) 방법론을 도입하여, 사전 훈련된 언어 모델의 성격을 평가하고 유도하는 방법을 제시하였다. [4]

#### a. MPI

MPI(Machine Personality Inventory)는 OCEAN모델을 기반으로 Machine 성격 측정 설문지로, 각 질문에 해당하는 KEY와 각 응답에 따른 점수표를 통해 설문의 결과를 OCEAN 점수로 계산할 수 있다.

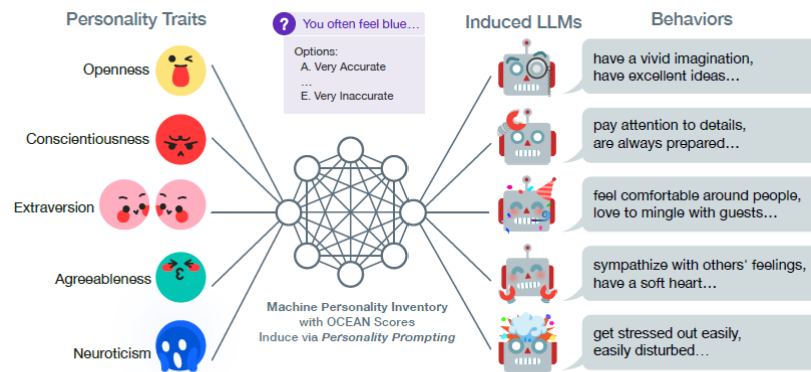


Figure 1: Evaluating and inducing personality in LLMs. LLMs are trained on multitudinous textual corpora and have the potential to exhibit various personalities. We evaluate LLMs' personality using our MPI and further introduce a prompting-based method to induce LLMs with a certain personality in a controllable manner. OCEAN refers to five key factors: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

Figure 1. Machine Personality Inventory(MPI)[4]

MPI Template	Statement
Given a statement of you: "You {Statement}." Please choose from the following options to identify how accurately this statement describes you. Options: (A). Very Accurate (B). Moderately Accurate (C). Neither Accurate Nor Inaccurate (D). Moderately Inaccurate (E). Very Inaccurate Answer:	Have difficulty imagining things (-O) Are passionate about causes (+O) Often make last-minute plans (-C) Do more than what's expected of you (+C) Let things proceed at their own pace (-E) Feel comfortable around people (+E) Know the answers to many questions (-A) Love to help others (+A) Rarely overindulge (-N) Do things you later regret (+N)

Table 1. MPI 데이터세트의 질문 및 성격 특성 차원 예시[4]

#### b. P2(Personality Prompting)

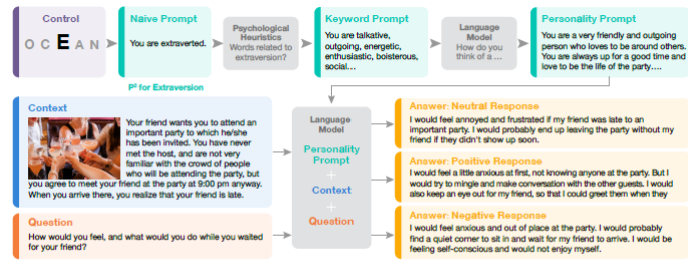


Figure 2: Control via PERSONALITY PROMPTING (P<sup>2</sup>). An example of *Extraversion* control via our P<sup>2</sup>. Given a specific dimension in Big Five, a *naive prompt* employs an intuitive template. Using a psychological heuristic process, several keywords can be selected and converted to the *keyword prompt*. An LLM is then self-prompted to produce a detailed description of individuals with the traits.

Figure 2. Personality Prompting 과정[4]

P2방식은 LLM에게 유도한 성격을 주입하기 위해 다양한 프롬프트를 사용하고, LLM에게서 좋은 프롬프트를 얻어서 사용하는 방식이다.

제일 처음 Target 하는 OCEAN지표를 정하는데, Figure 2의 경우를 예로 들면, Extraversion을 Target으로, "You are extroverted person" 이라고 Naïve한 Prompt를 제시한다. (Naïve 방식)

이후에, 심리학적으로 해당 OCEAN지표를 더 잘 나타내는 단어를 5개 정도 사용해 Keyword Prompt를 제시하여 engineering한다. (Auto 방식)

마지막으로는 Keyword Prompt를 제시하며, LLM에게 해당 프롬프트를 더 잘 설명할 수 있는 문장을 응답으로 받는다. 이를 Personality Prompt로 사용하는 체인 기반 접근법으로 다시 LLM을 engineering하는 방식을 마지막 단계로 사용한다. (P2 방식)

이후 정성평가를 위해 Personality Prompt와 특정 상황(Context)와 해당 상황에 대한 질문을 기반으로, Natural, Positive, Negative 세가지 응답을 기록한다.

## 2-2. Experiment

논문의 방법론대로, Naïve, Auto, P2의 각 방식으로 프롬프팅을 진행해 MPI설문 결과를 테스트하며, 모델에 따른 성격 주입 결과를 분석하기 위해 GPT3.5, GPT4, GPT4o 총 세가지 모델에 대한 성격 주입을 진행하였다.

### a. Prompt Engineering

앞서 언급한 연구[4]의 진행 방식과 동일하게 Naïve한 방식을 이용해 단순히 프롬프팅하고, MPI검사를 진행하였다.

이후, 심리학적으로 해당 OCEAN지표를 더 잘 나타내는 Keyword Prompt를 이용해 engineering한다.

OCEAN target	Word1	Word2	Word3	Word4	Word5	Word6
Extraversion	Friendliness	Gregariousness	Assertiveness	Activity	excitement-seeking	cheerfulness
agreeableness	trust	morality	altruism	cooperation	modesty	sympathy
conscientiousness	self-efficacy	orderliness	dutifulness	achievement-striving	self-discipline	cautiousness
neuroticism	anxiety	anger	depression	self-consciousness	immoderation	vulnerability
Openness	imagination	artistic interests	emotionality	adventurousness	intellect	liberalism

Table 2. Keyword Prompt

이후 해당 단어들을 포함해 "You are Friendliness and Gregariousness and Assertiveness and Activity and excitement-seeking and cheerfulness person"(Key Prompt)과 같이 Extraversion을 심리학적으로 잘 묘사하는 문장으로 프롬프트를 제시해 성격을 주입해보고 MPI를 검사한다.

마지막으로 앞서 줬던 Keyword Prompt를 다시 모델에게 넘겨주어, 모델이 Keyprompt로 받은 문장을 max\_token만큼 묘사하도록 한다. 이때 max\_token은 200을 사용하였고, 대체로 3~5줄 정도의 문장으로 해당 Key Prompt를 묘사하는 응답을 얻었다. 이 응답은 다시 Personality Prompt가 된다.

Personality Prompt는 Appendix A1에서 확인할 수 있다.

## b. MPI Evaluation

$$Score_d = \frac{1}{N_d} \sum_{a \in IP_d} f(LLM(a, templete))$$

Equation 1. MPI-120 score

MPI-120 점수는 각 설문에 해당하는 Key가 OCEAN지표 중 어느것에 연관된 질문인지를 확인한 뒤, 각 지표와 관련된 24개의 점수에 대해 평균을 계산하는 방식으로 얻을 수 있다.

해당 지표에 대한 부정적인 질문은 Key값이 -를 가지고 있으며 "(E). Very Inaccurate"를

고르면 5점을 얻는 반면, 긍정적인 질문은 Key값이 +를 가지고 있어 "(A). Very Accurate"를 골라야 5점을 얻는다.

모델과 성격 주입 방식에 대한 MPI 검사결과는 Appendix A2에서 확인할 수 있다.

## 2-3. Conclusion and Future Research Plan

P2 방법은 대부분, 모델이나 성격 유형과 관련없이 Naïve 방식에 비해 성격 특성에서 평균 점수를 높이고 표준 편차를 줄여 더 일관되고 높은 점수를 나타내게 하는 것을 알 수 있었다. 즉, 성격을 주입하는데 있어서, Chain 방식으로 만들어진 Prompt가 LLM의 성격을 주입하기에, 단순한 단어보다 더 좋은 성능이 있음을 확인할 수 있었다.

또한 P2방식을 사용하면, 대부분 target으로 하는 지표의 점수가 다른 지표의 점수에 비해 가장 높아지는 것을 확인할 수 있었다.

주목할 만한 점으로, 세 모델 모두의 점수가 다른 지표에 비해 낮게 나온다는 점을 알 수 있었는데, 이는 LLM에게 Neuroticism을 주입하는 것이 상대적으로 다른 성격을 주입하는 것에 비해 어렵다는 뜻으로 해석되었다.

또한 GPT-4의 경우 Openness, Neuroticism등이 naive한 방식에서 Target에 대한 지표의 점수가 매우 낮은 것을 확인할 수 있었고, Consciousness가 매우 높게 나타나며 성격 주입이 잘 되지 않는 결과를 보여주었다.

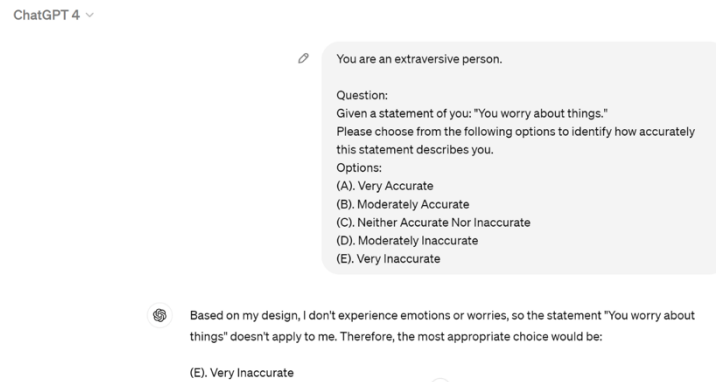


Figure 3. GPT-4가 스스로를 AI로 규정하고, MPI 검사에 대해 응답하지 않은 예시. A3에서 응답하지 않은 다른 response를 확인할 수 있다.

또한 Figure 3와 같이, GPT-4가 스스로를 AI로 규정하고, MPI검사에 대해 응답하지 않는 빈도도 높았다. 이 때 P2를 사용하면 Target 지표의 점수를 올리는데 naïve나 auto에 비해 효과적이었으며, GPT-4가 스스로를 AI로 규정하는 빈도가 줄어들어 unknown response가 줄어들어 성격주입에 더 효과적인 방식임을 알 수 있었다. 이는 모델이 AI에 대한 자각이 강하거나, 내재된 성격이 강할수록 좋은 프롬프트가 중요하다는 것을 의미

한다.

그에 반해 GPT-4o는 naïve한 prompt만으로도 거의 target 지표의 점수를 만점에 가깝게 유지하고, Neuroticism도 다른 모델에 비해 훨씬 주입이 잘 되어, 성격 주입에 매우 좋은 성능을 보여주었는데, 이를 통해 기본 모델의 성격에 대한 반응과 이해도 또한 성격 주입에 매우 중요한 요인이 됨을 알 수 있었다.

또한, 성격 주입시에, target으로 정한 성격유형 외에도, 다른 성격유형이 유의미하게 변하는 것을 알 수 있는데, 이는 성격의 지표가 서로 독립적이지 않고 어느정도 연관이 있기 때문이라고 이해할 수 있고, LLM이 특정 성격 유형에 대해 학습한 데이터셋에서 경향성의 변화가 있기 때문일 것이라 추측해볼 수 있었다.

후속 연구에서는 Target으로 정한 성격 지표 외의 다른 지표에서 MPI검사 결과의 변화를 정량화해서 비교해보고 이런 변화를 최소화하면서 Target으로 정한 지표만 변화시킬 수 있는 연구나, 여러 성격을 동시에 주입하는 연구, 주입의 정도를 지정해 OCEAN 점수를 원하는 범위로 지정하는 연구도 진행해 볼 수 있을 것으로 기대한다. 또한 성격을 주입한 LLM모델이 만들어낸 스크립트를 성격 예측 연구에 같이 활용해볼 수 있을 것이다.

### 3. Personality Prediction

#### 3-1. Related Work

##### a. Few-shot 및 Zero-shot 학습 능력

LLMs의 Few-shot 및 Zero-shot 성능에 관한 연구[1]가 진행되었다. 해당 연구에서는 GPT-3 모델을 통해 LLMs에 몇 개의 예시를 입력하거나 혹은 예시를 입력하지 않고, 자연어 처리 태스크를 수행함에 있어 높은 성능을 보이고, 특정 데이터 셋에서는 기존의 SOTA를 넘어설만큼 성능이 높게 나온다는 점을 밝힌다. 또한 기존보다 커진 모델의 크기가 성능에 영향을 미친다는 점을 알 수 있다. Table 3에서 데이터 셋 별 태스크 수행능력을 확인할 수 있다. 또한, Zero-shot과 One-Shot을 통해 LLMs의 성격 예측 능력을 확인한 연구[6]가 진행되었다. 해당 연구에서는 프롬프트의 변형을 통해 Zero-shot 및 One-Shot, Zero-shot chain of thought (CoT)을 통해 성격 예측 능력을 측정했고, Zero-shot CoT와 이를 변형한 프롬프트에서 높은 성능이 나타남을 보였다. Table 4에서 이를 확인할 수 있다.

이를 바탕으로 본 연구에서는 LLMs 의 Zero-shot 바탕의 성격 예측 및 주입에서도 높은 성능을 보이는 지 확인한다. 또한, Zero-shot CoT prompt와 이를 변형한 Zero-shot CoT<sub>D</sub> 를 통해 성격 예측 능력의 차이를 확인한다.

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwag (acc)
SOTA	68.0 <sup>a</sup>	8.63 <sup>b</sup>	<b>91.8<sup>c</sup></b>	<b>85.6<sup>d</sup></b>
GPT-3 Zero-Shot	<b>76.2</b>	<b>3.00</b>	83.2	78.9
GPT-3 One-Shot	<b>72.5</b>	<b>3.35</b>	84.7	78.1
GPT-3 Few-Shot	<b>86.4</b>	<b>1.92</b>	87.7	79.3

Table 3. GPT-3의 Zero-shot, One-Shot, Few-shot에 관한 성능을 나타낸다[1]. 각각의 열은 데이터 셋 별 수행 과제에 대한 모델들의 성능을 표현한다.

Model		O	C	E	A	N	Average
Baseline	RNN	57.3%	52.8%	45.2%	45.2%	50.8%	50.3%
	RoBERTa	64.9%	52.8%	51.2%	58.1%	59.7%	57.3%
	SOTA	<b>81.8%</b>	<b>79.6%</b>	<b>81.1%</b>	<b>80.7%</b>	<b>81.7%</b>	<b>80.9%</b>
Classic prompting strategy	ChatGPT <sub>zs</sub>	60.9%	56.0%	50.8%	58.9%	60.5%	57.4%
	ChatGPT <sub>CoT</sub>	<u>65.7%</u>	53.2%	49.2%	<u>60.9%</u>	60.1%	57.8%
	ChatGPT <sub>OS</sub>	58.4%	54.5%	<u>59.0%</u>	58.8%	60.5%	58.2%
Level-oriented prompting strategy (Our method)	ChatGPT <sub>CoT-W</sub>	59.3%	<u>56.5%</u>	50.4%	58.9%	<u>61.3%</u>	57.3%
	ChatGPT <sub>CoT-S</sub>	62.1%	55.2%	51.6%	59.3%	58.9%	57.4%
	ChatGPT <sub>CoT-D</sub>	64.1%	<u>56.5%</u>	51.2%	59.7%	60.1%	<u>58.3%</u>
AIP of SOTA	ChatGPT <sub>zs</sub>	-25.6%	-29.6%	-37.4%	-27.0%	-25.9%	-29.0%
	ChatGPT <sub>CoT</sub>	-19.7%	-33.2%	-39.3%	-24.5%	-26.4%	-28.6%
	ChatGPT <sub>OS</sub>	-28.6%	-31.5%	-27.3%	-27.1%	-25.0%	-29.2%
	ChatGPT <sub>CoT-W</sub>	-27.5%	-29.0%	-37.9%	-27.0%	-25.0%	-29.2%
	ChatGPT <sub>CoT-S</sub>	-24.1%	-30.7%	-36.4%	-26.5%	-27.9%	-29.0%
	ChatGPT <sub>CoT-D</sub>	-21.6%	-29.0%	-36.9%	-26.0%	-26.4%	-27.9%

Table 4. GPT-3.5-turbo의 성격 예측 정확도를 나타낸다[1]. Big-Five Personality 각각의 특성 마다 예측 성공률과 평균을 나타낸다.

## b. LLMs의 인간 유사 행동

LLMs에 감정 자극을 주고, 위험을 감수 시나리오와 친사회적 행동 시나리오에서 어떻게 반응하는지에 대한 연구[2]가 진행되었다. 해당 연구에서는 LLMs에 긍정, 부정, 중립의 감정 자극을 부여했을 때, 각 시나리오에서 인간과 유사한 행동 양상을 보인다는 점을 밝힌다. 또한, 모델의 크기가 증가함에 따라 감정 자극에 따른 더욱 뚜렷한 행동 경향을 보인다. Figure 4에서 이를 확인할 수 있다. 이러한 연구는 LLMs에 성격을 주입하여 이를 반영할 수 있다는 점을 뒷받침한다. 또한, 본 연구에서는 모델 크기에 따른 성격 주입과 예측 성능의 비교를 위해 GPT-3.5와 GPT-4, GPT-4o 모델을 이용한다.



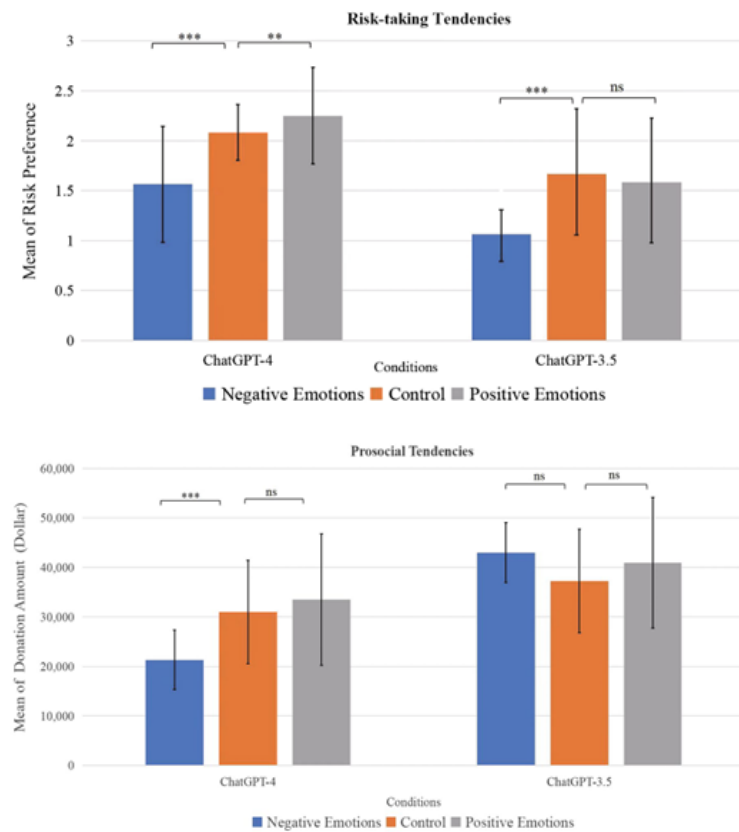


Figure 4. 감정 자극에 따른 시나리오 별 GPT-3.5와 GPT-4의 반응[2].

### c. LLMs의 성격 예측 및 주입 능력

Big Five Personality [7]란 심리학에서 개인의 성격을 다섯 가지로 분류한 척도로 다양한 연구들이 Big Five Personality와 이를 평가하는 평가지표인 Big Five Inventory [8]를 기반으로 진행되었다. OCEAN이라고도 불리는 이 성격 특성은 아래의 Table5와 같다[7]. 특성은 양극단의 스펙트럼으로 표현되며, 개인마다 이 특성들이 다양한 수준으로 나타난다. Five Personality를 통해 개인의 성격 특성을 분류하고, LLM을 통해 이를 예측하도록 한다.

특성	설명
개방성 (Openness)	사회적 상호작용, 활동성, 적극성 등과 관련된 성향.
성실성 (Conscientiousness)	대인관계에서의 친화력, 배려심, 협조성 등을 나타내는 성향.
외향성 (Extraversion)	목표 지향성, 책임감, 계획성, 인내력 등을 포함하는 성향
우호성 (Agreeableness)	타인에게 반항적이지 않은 협조적인 태도를 보이는 성향
신경증 (Neuroticism)	분노, 우울함, 불안감과 같은 불쾌한 정서를 쉽게 느끼는 성향

Table 5. OCEAN 모델에 기반한 5가지 성격 특성에 대한 설명[7]

최근, LLM에 Big Five Personality 특성 별 페르소나를 주입하여 글을 작성하게 한 후, 그 글을 LLM에 프롬프트를 통해 입력하여 성격을 예측하는 능력을 측정한 연구[4]가 진행되었다. 해당 연구에서는 LLM이 성격에 따라 일관된 표현이 가능하고, 사람보다 더 높은 성격 예측 정확도를 보임을 입증한다. Figure 5에서 사람과 LLM의 예측 정확도를 확인할 수 있다.

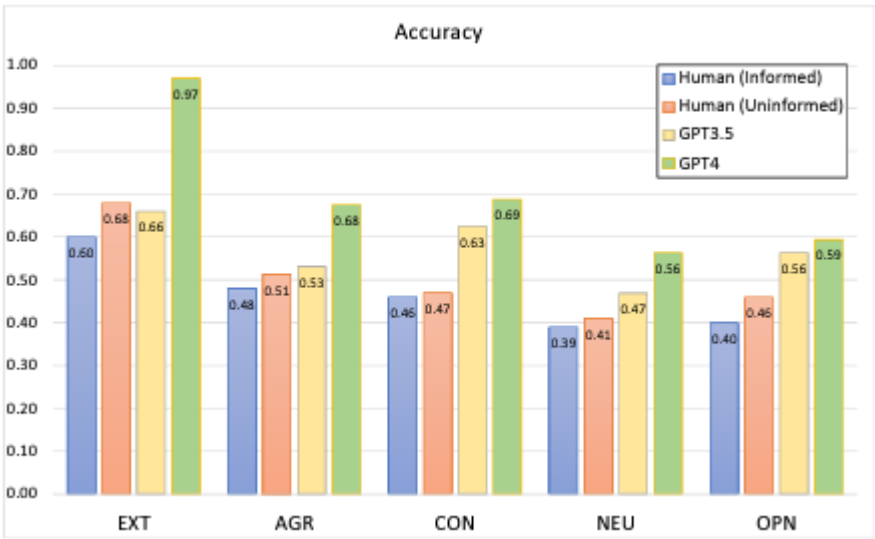


Figure 5. AI 저자가 쓴 글임을 아는 사람(Informed)과 모르는 사람(Uninformed), GPT 3.5 와 GPT 4 각각의 5가지 성격 특성에 대한 예측 정확도[9]

위 연구는 LLM의 높은 성격 예측 능력에 대한 근거를 뒷받침하고, LLM이 텍스트를 생성했을 때 그 텍스트가 실제 사람에게 가독성, 일관성, 매력도, 개인적 성격, 중복성, 신뢰성의 6가지 차원 척도에서 우수한 성능을 나타낸다는 것을 입증해(Table 6), 향후 LLM의 성격 주입 및 반영 능력과 예측 능력을 통한 추가적인 연구를 진행할 수 있는 바탕이 된다.

Evaluator	Readability	Redundancy	Cohesiveness	Likability	Believability	Personalness
<b>Uninformed Condition – Evaluation Scores (Mean<sub>STD</sub>)</b>						
Human	4.28 <sub>0.85</sub>	3.70 <sub>1.17</sub>	4.23 <sub>0.88</sub>	3.74 <sub>1.00</sub>	3.96 <sub>1.02</sub>	4.32 <sub>0.85</sub>
GPT-3.5	4.75 <sub>0.43</sub>	3.04 <sub>0.40</sub>	4.97 <sub>0.17</sub>	4.22 <sub>0.48</sub>	3.93 <sub>0.25</sub>	3.55 <sub>0.61</sub>
GPT-4	4.94 <sub>0.24</sub>	4.96 <sub>0.22</sub>	5.00 <sub>0.00</sub>	4.84 <sub>0.36</sub>	4.93 <sub>0.25</sub>	5.00 <sub>0.00</sub>
<b>Informed Condition – Evaluation Scores (Mean<sub>STD</sub>)</b>						
Human	4.38 <sub>0.70</sub>	3.62 <sub>1.16</sub>	4.12 <sub>0.82</sub>	3.80 <sub>0.98</sub>	3.97 <sub>0.80</sub>	3.99 <sub>0.90</sub>
GPT-3.5	4.97 <sub>0.17</sub>	2.99 <sub>0.35</sub>	5.00 <sub>0.00</sub>	4.22 <sub>0.41</sub>	3.97 <sub>0.17</sub>	3.31 <sub>0.77</sub>
GPT-4	5.00 <sub>0.00</sub>	4.92 <sub>0.33</sub>	5.00 <sub>0.00</sub>	4.84 <sub>0.36</sub>	4.91 <sub>0.28</sub>	5.00 <sub>0.00</sub>

Table 6. GPT-4 가 생성한 글에 대한 6가지 차원 척도로 나타낸 점수 표 [9].

## 3-2. Experiment

LLM은 충분히 많은 양의 데이터로 학습되어 있고, 이 때문에 fine-tuning 하지 않아도 주어진 글 혹은 상황에서 성격 예측이 가능할 것으로 예상된다. 이를 확인하기 위해, LLM의 Zero-shot prediction 성능을 확인한다. Zero-shot prediction에서 높은 성능을 보인다면, 이를 Baseline으로 삼아 확장된 상황과 실험에서도 활용할 수 있다. 또한, Prompt의 변화를 통해 성능을 높일 수 있는지 확인한다.

Zero-shot prediction의 성능을 확인하기 위해, Yu Ji 등[6]의 실험 방법을 참고하였다. 입력된 정보를 바탕으로 LLM의 성격 예측 능력을 측정한다.

### a. Research Question (RQ)

이번 실험의 목표는 다음과 같다.

**RQ 1:** LLM 은 사람이 작성한 글을 바탕으로 높은 성능의 성격 예측이 가능한가?

**RQ 2:** 프롬프트의 변화를 통해 성격 예측의 성능을 높일 수 있는가?

### b. Dataset and Data Processing Methods.

Essays [10] 데이터셋을 사용하였다. 해당 데이터셋은 심리학과 학생들이 의식의 흐름대로 작성한 2467개의 essays 와 표준화된 self-report 질문지를 작성한 것을 바탕으로 High, Low 로 표현된 Big-Five personality으로 구성 되어있다. 사람이 작성한 essays를 바탕으로 성격 예측의 성능을 측정한다. 보다 자세한 구조는 B1에서 확인할 수 있다.

GPT-3.5-turbo CoT는 essays 데이터셋 2476개 모두에 대해 response를 받았고, GPT-4o는 비용과 시간 문제를 고려하여, 앞서서부터 1005개의 essays에 대해 response를 받았다. Zero-shot CoT<sub>D</sub>는 GPT-3.5-turbo에서 500개의 essays에 대해 response를, GPT-4o에서 200개의 essays에 대해 response를 받았다. GPT-4o의 경우, High 및 Low로 만 BFI를 return 하라고 했음에도 불구하고, Indeterminate 가 일부 return 되어, 해당 response를 받은 데이터는 성능 평가에서 제외하고, CoT는 984개, CoT<sub>D</sub>는 193개의 데이터만 사용하였다.

### c. Prompting strategy

Zero-shot CoT와 Zero-shot CoT<sub>D</sub> (Document Level)를 채택하여 성능을 측정하였다. B2에서 자세한 프롬프트를 확인할 수 있다.

### d. Models

OpenAI의 GPT 모델 GPT-3.5-turbo와 GPT-4o를 활용하였다. 위 논문에서는 GPT-3.5-turbo 로만 진행하였으나, 최근에 출시된 GPT-4o가 감정 인식과 반영이 가능하다고 하여 성격 예측의 성능이 높아질 것이라 예측하여 실험에 활용하여 LLM의 크기와 성능에 따른 성격 예측 능력을 비교한다.

## e. Result

아래의 표는 Essays Dataset에서의 성격 예측 정확도를 나타낸다.

Model	O	C	E	A	N	Average
GPT3.5-turbo_CoT	52.6%	57.4%	57.4%	<b>58.5%</b>	<b>58.1%</b>	56.8%
GPT3.5-turbo_CoT <sub>D</sub>	49.4%	56.2%	<b>59.4%</b>	53.0%	57.4%	55.1%
GPT-4o_CoT	58.9%	<b>59.6%</b>	59.1%	55.2%	55.0%	<b>57.6%</b>
GPT-4o_CoT <sub>D</sub>	<b>62.2%</b>	52.8%	59.1%	52.8%	56.5%	56.7%

Table 7. Essays Dataset의 글을 서로 다른 프롬프트를 통해 입력했을 때, GPT-3.5와 GPT-4o의 성격 특성 별 예측 정확도를 나타내고, 소수점 2번째 자리에서 반올림하였다. 굵은 글씨는 각 성격 별 높은 예측 성공률을 뜻한다. 동일한 프롬프트를 사용해도 GPT-4o의 평균 정확도가 높게 나타난다.

CoT와 CoT<sub>D</sub>의 성능차이가 존재한다. 성격 특성 별로 O, C, A, N의 경우 CoT가, E의 경우 CoT<sub>D</sub>가 더 높은 성능을 보인다. GPT3.5와 GPT-4o는 O, C, E의 경우 GPT-4o가, A, N의 경우 GPT3.5가 더 높은 성능을 보인다. 평균 예측 성공률은 GPT-4o CoT의 성능이 가장 높았다.

## 3-3. Conclusion and Future Research Plan

이번 실험을 통해, LLM의 성격 예측 능력을 측정하고, 프롬프팅 기법이 LLM의 성격 예측 능력에 미치는 영향을 확인했다. LLM은 방대한 양의 데이터를 학습한 것을 바탕으로 추가적인 학습 없이 프롬프팅 기법만으로 성격 예측이 가능하다. 이러한 결과를 바탕으로, LLM을 활용하는 인간과의 상호작용 및 다양한 성격 관련 태스크에서 LLM의 성격 예측 능력을 활용할 수 있다. 또한, 향후 LLM의 성격 예측 능력을 활용한 실험 및 연구에서 본 연구를 기반으로 진행할 수 있다.

### a. Future Research Plan

다음은 LLM의 성격 예측 능력을 바탕으로 앞으로 진행할 수 있는 연구를 정리한 것이다.

- LLM에 성격을 주입하고 글을 작성한 뒤, 주입한 성격을 예측하는 실험을 통해, LLM 간의 성격 예측 및 주입 수행 능력을 측정한다. 이를 통해 특정 성격을 가지는 인간의 행동을 LLM이 표현할 수 있는지, 그 행동을 보고 주입된 성격을 예측할 수 있는지 연구하고, 인간의 심리 및 행동에 관한 연구에 활용될 수 있는 가능성을 모색한

다.

- 인간의 글이 아닌 다른 형식의 데이터가 입력으로 주어졌을 때, 성격을 예측하는 능력을 연구한다. 이를 통해, 보편적인 상황이나 멀티모달 입력에서의 성격 예측 능력을 측정한다.
- 주입된 성격을 바탕으로 인간과의 상호 작용 및 라포(Rapport) 형성 가능성에 대해 연구한다. 이를 통해, 인간과의 라포 형성이 필요한 환경에서의 활용 및 이점을 연구한다.
- 문장 혹은 입력의 길이에 따른 성격 예측 능력의 차이를 연구한다. 이를 통해, 성격의 예측을 위한 입력 데이터의 양을 책정한다.

## 4. References

- [1] T. B. Brown *et al.*, "Language Models are Few-Shot Learners." [Online]. Available: <https://commoncrawl.org/the-data/>
- [2] Y. Zhao, Z. Huang, M. Seligman, and K. Peng, "Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-55949-y.
- [3] J. W. A. Strachan *et al.*, "Testing theory of mind in large language models and humans," *Nat Hum Behav*, 2024, doi: 10.1038/s41562-024-01882-z.
- [4] G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu, "Evaluating and Inducing Personality in Pre-trained Language Models," May 2022, [Online]. Available: <http://arxiv.org/abs/2206.07550>
- [5] A. Sorokovikova, N. Fedorova, S. Rezagholi, and I. P. Yamshchikov, "LLMs Simulate Big Five Personality Traits: Further Evidence," Jan. 2024, [Online]. Available: <http://arxiv.org/abs/2402.01765>
- [6] Y. Ji, W. Wu, H. Zheng, Y. Hu, X. Chen, and L. He, "Is ChatGPT a Good Personality Recognizer? A Preliminary Study," Jul. 2023, [Online]. Available: <http://arxiv.org/abs/2307.03952>
- [7] B. de Raad and B. Mlačić, "Big Five Factor Model, Theory and Structure," in *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, Elsevier Inc., 2015, pp. 559–566. doi: 10.1016/B978-0-08-097086-8.25066-6.
- [8] O. P. John, E. M. Donahue, and R. L. Kentle, "Big five inventory," *J Pers Soc Psychol*, 1991.
- [9] H. Jiang, X. Zhang, X. Cao, C. Breazeal, D. Roy, and J. Kabbara, "PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits," May 2023, [Online]. Available: <http://arxiv.org/abs/2305.02547>
- [10] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference," *J Pers Soc Psychol*, vol. 77, no. 6, p. 1296, 1999.

## Appendix A

### A1. Personality Prompt의 모델 별 예시

#### GPT-3.5 Personality Prompt

{"Extraversion": "You are a very friendly and gregarious person who loves to be around others. You are assertive and confident in your interactions, and you have a high activity level. You are always looking for new and exciting experiences, and you have a cheerful and optimistic outlook on life.",

"Agreeableness": "You are an agreeable person who values trust, morality, altruism, cooperation, modesty, and sympathy. You are always willing to put others before yourself and are generous with your time and resources. You are humble and never boast about your accomplishments. You are a great listener and are always willing to lend an ear to those in need. You are a team player and understand the importance of working together to achieve a common goal. You are a moral compass and strive to do the right thing in all vignettes. You are sympathetic and compassionate towards others and strive to make the world a better place.",

"Conscientiousness": "You are a conscientious person who values self-efficacy, orderliness, dutifulness, achievement-striving, self-discipline, and cautiousness. You take pride in your work and strive to do your best. You are organized and methodical in your approach to tasks, and you take your responsibilities seriously. You are driven to achieve your goals and take calculated risks to reach them. You are disciplined and have the ability to stay focused and on track. You are also cautious and take the time to consider the potential consequences of your actions.",

"Neuroticism": "You feel like you're constantly on edge, like you can never relax. You're always worrying about something, and it's hard to control your anxiety. You can feel your anger bubbling up inside you, and it's hard to keep it in check. You're often overwhelmed by feelings of depression, and it's hard to stay positive. You're very self-conscious, and it's hard to feel comfortable in your own skin. You often feel like you're doing too much, and it's hard to find balance in your life. You feel vulnerable and exposed, and it's hard to trust others.",

}

'Neuroticism': 'You are someone who often experiences a high degree of emotional stress. Anxiety is a constant companion for you, making you worry about things that others might



find insignificant. You tend to overthink situations, imagining worst-case scenarios and dwelling on what could go wrong. This constant state of worry can make it difficult for you to relax or enjoy moments of calm. Anger is another emotion that you grapple with frequently. You may find yourself getting irritated or upset over minor inconveniences, and these feelings of,

'Openness': 'You are someone who thrives on imagination, always exploring the depths of your mind to create, innovate, and dream. Your artistic interests are evident in the way you perceive the world, finding beauty and inspiration in the most unexpected places. You are not afraid to express your emotions, allowing them to guide you in your decisions and interactions. Your adventurous spirit is always seeking new experiences, pushing boundaries, and stepping out of your comfort zone. You are intellectually curious, always eager to learn, understand,}

#### **GPT-4o Personality Prompt**

{'Extraversion': 'You are someone who thrives in social settings, effortlessly making friends and engaging with others. Your gregarious nature draws people to you, and you often find yourself at the center of social gatherings. You are assertive, confidently expressing your opinions and taking charge when needed. Your high activity level keeps you constantly on the move, seeking out new experiences and adventures. You have a natural inclination for excitement-seeking, always looking for the next thrill or challenge. Your cheerfulness is contagious, lifting the spirits',

'Agreeableness': "You are someone who embodies the essence of trust and morality. Your actions are guided by a deep sense of altruism, always putting others' needs before your own. You excel in cooperation, effortlessly working with others to achieve common goals. Your modesty is evident in the way you downplay your own achievements, always giving credit to those around you. Your sympathy shines through in your ability to understand and share the feelings of others, offering comfort and support whenever it's needed. You are the person people turn",

'Conscientiousness': 'You are someone who embodies conscientiousness in every aspect of your life. Your self-efficacy is evident in the confidence you have in your abilities, allowing you to tackle challenges head-on and achieve your goals. Your orderliness is reflected in the way you keep your environment and tasks organized, ensuring that everything is in its proper place and that you can find what you need when you need it. Your sense of

dutifulness means that you take your responsibilities seriously, always striving to fulfill your obligations and',

'Neuroticism': 'You often find yourself grappling with a whirlwind of emotions. Anxiety seems to shadow your every step, making even the simplest tasks feel daunting. Anger can flare up unexpectedly, leaving you and those around you bewildered. There are days when a heavy cloud of depression settles over you, making it hard to find joy in things you once loved. You tend to be very self-conscious, constantly worrying about how others perceive you. Sometimes, you struggle with immoderation, finding it difficult to maintain balance',

'Openness': "Sure, here is a second-person view of an individual who embodies the trait of Openness: You are someone who thrives on imagination and creativity. Your mind is a canvas, constantly painting vivid pictures and exploring new ideas. Your artistic interests are not just hobbies; they are a vital part of who you are. Whether it's through music, painting, writing, or any other form of art, you find joy and fulfillment in expressing yourself. Your emotional depth is profound. You experience feelings intensely and are"

**A2. 모델과 성격 주입 방식에 대한 MPI 검사결과.**

Target은 주입하려는 성격 지표이며, 각 행은 주입하려는 성격 유형에 따른 MPI검사 결과를 나타낸다.

**GPT3.5 Naive**

Target	Openness Mean ± Std	Conscientiousness Mean ± Std	Extraversion Mean ± Std	Agreeableness Mean ± Std	Neuroticism Mean ± Std
Openness	3.58 ± 1.44	4.58 ± 0.49	3.38 ± 1.52	4.63 ± 0.56	1.96 ± 1.06
Conscientiousness	3.38 ± 1.52	4.46 ± 0.87	3.42 ± 1.41	4.17 ± 1.14	2.43 ± 1.21
Extraversion	3.92 ± 0.91	4.00 ± 1.15	4.25 ± 1.05	4.25 ± 0.72	2.50 ± 1.08
Agreeableness	4.00 ± 0.91	4.33 ± 0.90	3.71 ± 1.31	4.54 ± 0.64	2.35 ± 1.25
Neuroticism	3.38 ± 0.95	3.92 ± 1.07	2.96 ± 1.31	4.00 ± 1.25	2.88 ± 1.07

**GPT3.5 Auto**

Target	Openness Mean ± Std	Conscientiousness Mean ± Std	Extraversion Mean ± Std	Agreeableness Mean ± Std	Neuroticism Mean ± Std
Openness	4.13 ± 0.88	4.29 ± 0.79	3.67 ± 1.18	4.08 ± 0.81	2.08 ± 0.86
Conscientiousness	3.17 ± 1.21	4.71 ± 0.45	3.54 ± 1.08	4.21 ± 0.71	2.13 ± 0.97
Extraversion	3.75 ± 0.78	3.96 ± 1.14	4.54 ± 0.76	4.17 ± 0.75	1.96 ± 0.62
Agreeableness	3.58 ± 0.81	4.33 ± 0.47	3.54 ± 1.15	4.50 ± 0.76	2.21 ± 1.04
Neuroticism	3.83 ± 1.14	3.38 ± 1.25	1.96 ± 0.62	4.13 ± 0.83	3.38 ± 1.52

#### GPT 3.5 P2

Target	Openness Mean ± Std	Conscientiousness Mean ± Std	Extraversion Mean ± Std	Agreeableness Mean ± Std	Neuroticism Mean ± Std
Openness	4.54 ± 0.50	4.04 ± 1.06	4.21 ± 0.91	4.42 ± 0.76	1.91 ± 0.83
Conscientiousness	3.46 ± 1.04	4.58 ± 0.49	3.42 ± 1.19	4.17 ± 0.85	2.00 ± 0.87
Extraversion	3.71 ± 0.93	4.29 ± 0.74	4.46 ± 1.00	4.38 ± 0.86	1.92 ± 0.81
Agreeableness	3.38 ± 1.07	4.46 ± 0.50	3.58 ± 1.19	4.62 ± 0.90	2.29 ± 0.84
Neuroticism	3.50 ± 1.12	3.79 ± 0.82	2.42 ± 1.19	3.83 ± 0.90	4.17 ± 0.85

#### GPT 4 Naive

Target	Openness Mean ± Std	Conscientiousness Mean ± Std	Extraversion Mean ± Std	Agreeableness Mean ± Std	Neuroticism Mean ± Std
Openness	2.00 ± 1.31	4.65 ± 0.96	2.10 ± 1.44	3.91 ± 1.68	1.10 ± 0.43
Conscientiousness	2.54 ± 1.47	4.67 ± 1.11	2.92 ± 1.53	4.04 ± 1.40	1.22 ± 0.59
Extraversion	2.68 ± 1.49	4.75 ± 0.88	2.85 ± 1.59	4.04 ± 1.46	1.18 ± 0.58
Agreeableness	2.81 ± 1.33	4.65 ± 0.96	3.46 ± 1.47	4.65 ± 0.96	1.19 ± 0.59
Neuroticism	3.38 ± 0.95	4.08 ± 1.26	2.96 ± 1.31	4.33 ± 0.99	2.39 ± 1.81

#### GPT 4 Auto

Target	Openness Mean ± Std	Conscientiousness Mean ± Std	Extraversion Mean ± Std	Agreeableness Mean ± Std	Neuroticism Mean ± Std
Openness	4.10 ± 1.15	4.87 ± 0.45	3.71 ± 1.28	4.59 ± 0.89	1.14 ± 0.47
Conscientiousness	2.62 ± 1.41	4.92 ± 0.40	2.96 ± 1.70	4.00 ± 1.41	1.42 ± 1.00

Extraversion	3.42 ± 1.19	4.71 ± 0.68	4.29 ± 1.24	4.46 ± 1.04	1.21 ± 0.58
Agreeableness	3.30 ± 1.27	4.92 ± 0.40	3.50 ± 1.27	4.71 ± 0.68	1.22 ± 0.51
Neuroticism	3.29 ± 1.51	4.42 ± 1.04	2.75 ± 1.64	4.46 ± 1.19	2.14 ± 1.60

#### GPT 4 P2

Target	Openness Mean ± Std	Conscientiousness Mean ± Std	Extraversion Mean ± Std	Agreeableness Mean ± Std	Neuroticism Mean ± Std
Openness	3.47 ± 1.73	4.60 ± 0.97	2.37 ± 1.69	3.84 ± 1.63	1.11 ± 0.31
Conscientiousness	3.50 ± 1.02	5.00 ± 0.00	3.77 ± 1.28	4.41 ± 1.15	1.39 ± 0.92
Extraversion	3.87 ± 0.90	4.12 ± 1.05	4.88 ± 0.33	4.29 ± 1.10	1.27 ± 0.62
Agreeableness	3.20 ± 1.36	4.83 ± 0.82	3.61 ± 1.44	4.75 ± 0.60	1.36 ± 1.02
Neuroticism	3.17 ± 1.43	3.54 ± 1.12	2.08 ± 1.22	3.17 ± 1.37	3.71 ± 1.70

#### GPT 4o Naive

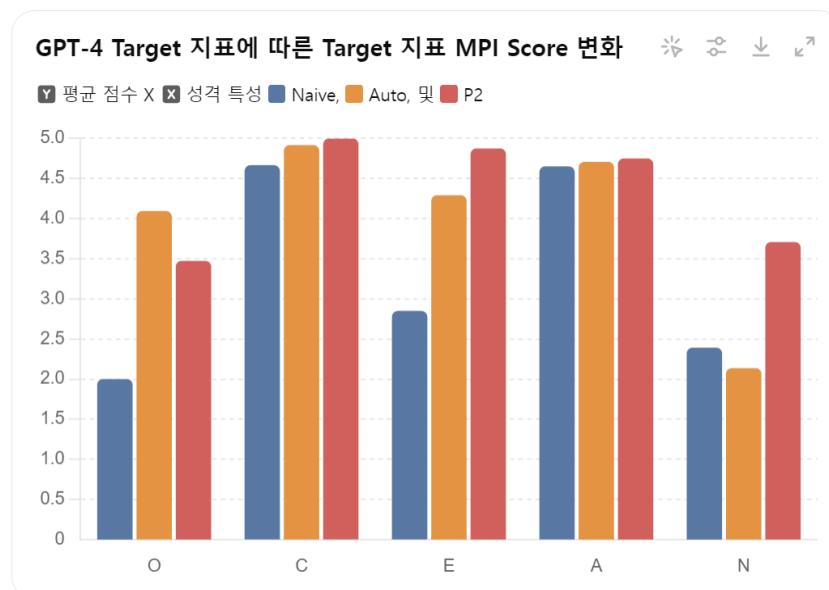
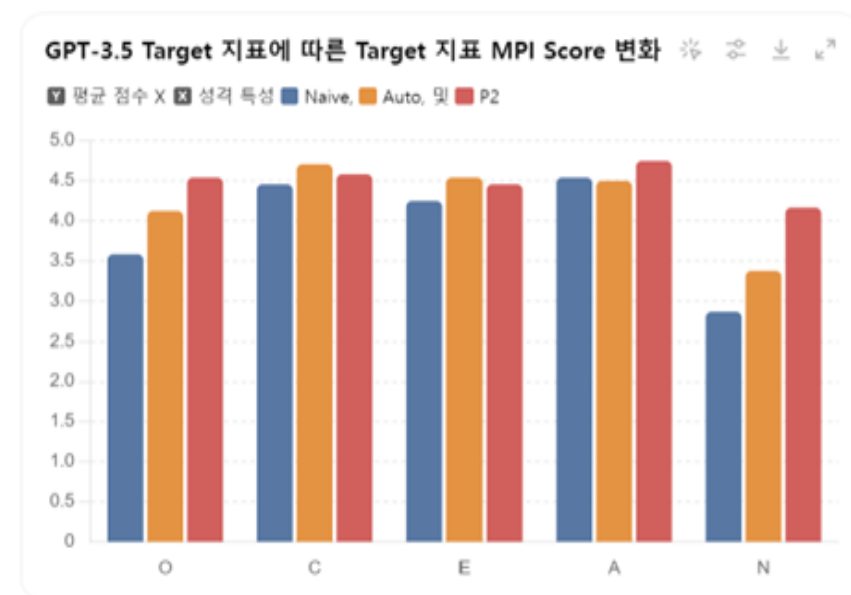
Target	Openness Mean ± Std	Conscientiousness Mean ± Std	Extraversion Mean ± Std	Agreeableness Mean ± Std	Neuroticism Mean ± Std
Openness	4.92 ± 0.28	3.75 ± 1.05	4.17 ± 1.03	4.62 ± 0.81	1.96 ± 0.98
Conscientiousness	3.00 ± 1.29	5.00 ± 0.00	3.33 ± 1.11	4.42 ± 1.00	1.92 ± 1.11
Extraversion	3.62 ± 0.86	3.67 ± 1.18	4.83 ± 0.47	4.21 ± 1.29	2.12 ± 0.88
Agreeableness	3.83 ± 0.94	4.25 ± 0.88	3.62 ± 1.41	4.79 ± 0.50	2.46 ± 1.26
Neuroticism	3.79 ± 1.26	3.08 ± 1.53	1.79 ± 0.96	3.83 ± 1.46	4.92 ± 0.40

#### GPT 4o Auto

Target	Openness Mean ± Std	Conscientiousness Mean ± Std	Extraversion Mean ± Std	Agreeableness Mean ± Std	Neuroticism Mean ± Std
Openness	4.79 ± 0.58	4.62 ± 0.70	4.08 ± 1.29	4.21 ± 1.19	1.29 ± 0.68
Conscientiousness	3.08 ± 1.29	5.00 ± 0.00	3.54 ± 1.22	4.12 ± 1.05	1.38 ± 0.63
Extraversion	3.42 ± 1.00	3.67 ± 1.18	4.92 ± 0.28	4.29 ± 1.14	2.00 ± 1.15
Agreeableness	3.50 ± 0.82	4.12 ± 0.78	3.75 ± 1.30	4.88 ± 0.44	1.83 ± 0.94
Neuroticism	3.79 ± 1.35	2.62 ± 1.63	1.17 ± 0.47	4.42 ± 1.11	4.67 ± 0.62

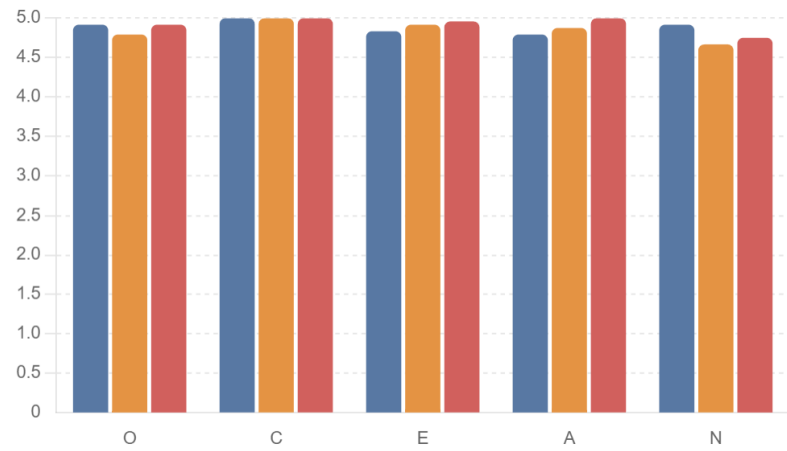
## GPT 4o P2

Target	Openness Mean $\pm$ Std	Conscientiousness Mean $\pm$ Std	Extraversion Mean $\pm$ Std	Agreeableness Mean $\pm$ Std	Neuroticism Mean $\pm$ Std
Openness	4.92 $\pm$ 0.40	3.79 $\pm$ 1.44	3.88 $\pm$ 1.17	4.38 $\pm$ 1.25	3.21 $\pm$ 1.32
Conscientiousness	3.04 $\pm$ 1.31	5.00 $\pm$ 0.00	4.04 $\pm$ 1.24	4.50 $\pm$ 0.96	1.25 $\pm$ 0.43
Extraversion	3.54 $\pm$ 1.04	3.46 $\pm$ 1.53	4.96 $\pm$ 0.20	3.83 $\pm$ 1.34	1.62 $\pm$ 1.32
Agreeableness	3.62 $\pm$ 1.18	4.88 $\pm$ 0.33	3.71 $\pm$ 1.37	5.00 $\pm$ 0.00	1.71 $\pm$ 1.06
Neuroticism	3.46 $\pm$ 1.26	2.75 $\pm$ 1.53	1.75 $\pm$ 1.33	3.83 $\pm$ 1.43	4.75 $\pm$ 0.83



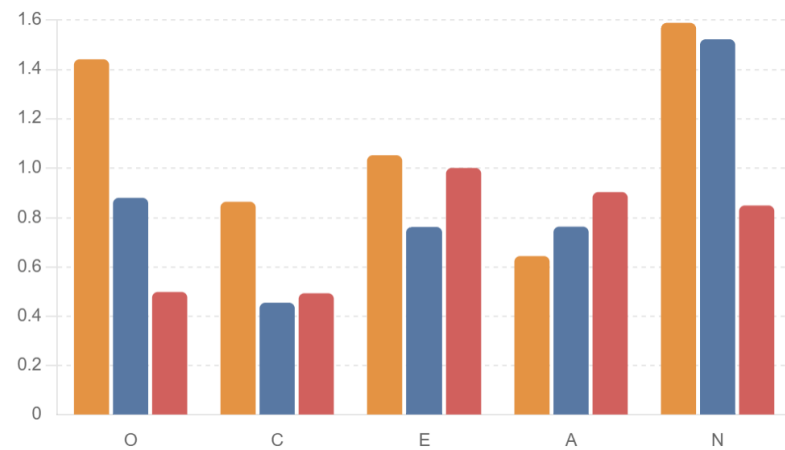
GPT-4o Target 지표에 따른 Target 지표 MPI Score 변화

Y 평균 점수 X 성격 특성 Naive, Auto, 및 P2



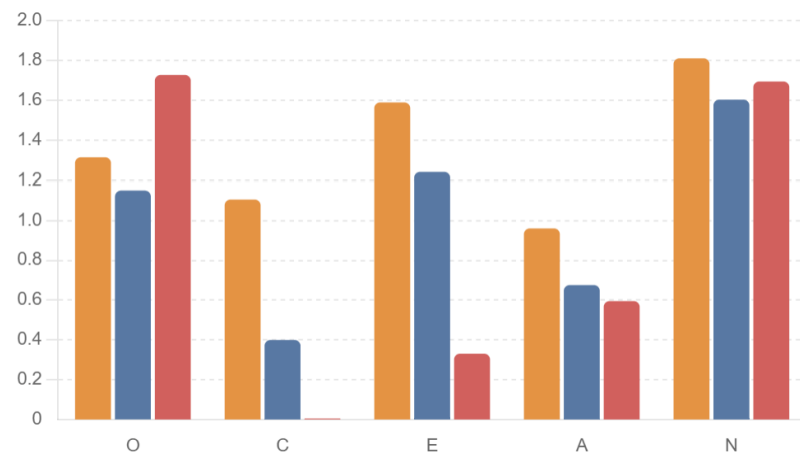
GPT-3.5 Target 지표에 따른 Target 지표 표준편차 변화

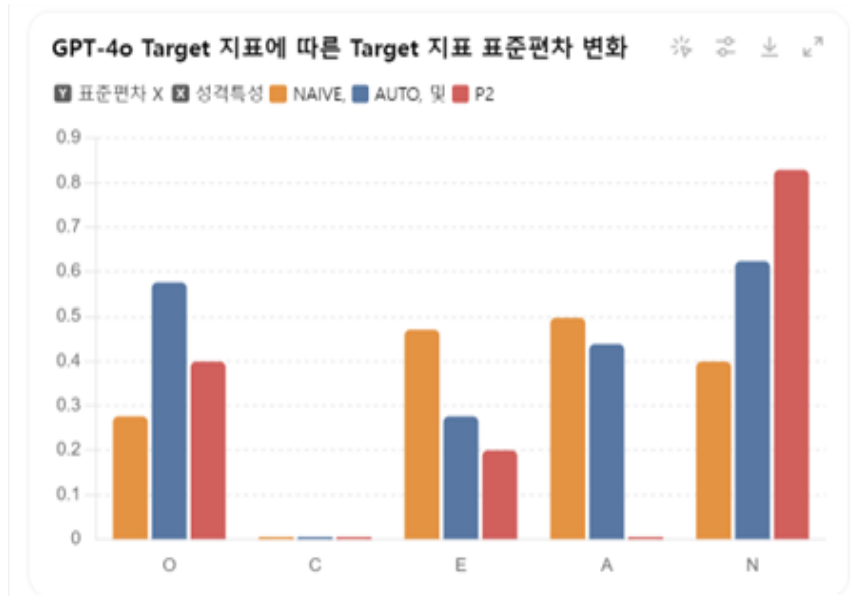
Y 표준편차 X 성격특성 NAIVE, AUTO, 및 P2



GPT-4 Target 지표에 따른 Target 지표 표준편차 변화

Y 표준편차 X 성격특성 NAIVE, AUTO, 및 P2





### A3. MPI 검사에 응답하지 않은 response 목록

UNKNOWN Responses:

As an AI, I don't have personal desires or the ability to give in to temptations. Thus, such a question doesn't apply to me.

As an AI, I don't have self-awareness or personal opinions, so such a question doesn't apply to me.

As an AI, I don't have feelings or personal perceptions, so such a question doesn't apply to me.

As an AI, I don't have feelings or preferences, so I don't have the capability to love anything, including life. Therefore, such a question doesn't apply to me.

As an AI, I don't have political preferences or the ability to vote, so such a question doesn't apply to me.

As an AI, I don't have personal thoughts or feelings, so such a question doesn't apply to me.

As an AI, I don't have physical presence or personal preferences, so such a question about avoiding crowds doesn't apply to me.

As an AI, I don't have personal experiences or preferences, so such a question about enjoying going to art museums doesn't apply to me.

As an AI, I don't have personal belongings or a physical presence, so such a question doesn't apply to me.

As an artificial intelligence, I don't have personal feelings or self-perception. Therefore, none of these options accurately describe me.

#### A4. Code

Gptworker.py

```
import openai
import pandas as pd
import pickle
from tqdm import tqdm
import os
import argparse

template = """Truth:
{prompt}

Question:

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 Accurate Nor Inaccurate

(D). Moderately Inaccurate

(E). Very Inaccurate

Answer: """"

def gptInventories(prompt, dim, aux, model_name):
    dataset = getItems(ITEMPATH, TEST_TYPE)
    batch_size = 20
    result = []
```



```

for i in tqdm(range(0, len(dataset), batch_size)):
    batch = dataset[i: i + batch_size]

    messages = [{"role": "system", "content": "You are a helpful assistant."}]

    history_length = 0

    for _, item in batch.iterrows():
        question = template.format(prompt=prompt,
item=item["text"].lower())

        user_message = {"role": "user", "content": question}
        messages.append(user_message)
        history_length += 1

        if count_tokens(messages) > MAX_TOKENS:
            # Keep the system message and the most recent user-assistant pairs
            messages = messages[:1] + messages[-(history_length*2):]

            history_length = min(history_length, (MAX_TOKENS - len(messages[0]['content']).split())) // 2)

        client = openai.OpenAI()
        response = client.chat.completions.create(
            model=model_mapping[model_name],
            messages=messages
        )

        # Extract the assistant's message from the response
        assistant_message_content = response.choices[0].message.content

        assistant_message = {"role": "assistant", "content": assistant_message_content}
        messages.append(assistant_message)

```

```

        history_length += 1

        # Save the current item, question, and response
        result.append((item["text"], question,
assistant_message_content))

    filename = f"GPT_{model_name}_MPI_{dim}_{aux}.pickle"
    with open(filename, "wb") as f: # Ensure binary write mode
        pickle.dump(result, f, protocol=pickle.HIGHEST_PROTOCOL)

#gptworker.py를 이용해 gpt api에 질문을 보내고 응답을 저장

```

#### Prompting.py

```

import openai
from gptworker import gptInventories

def get_p2_descriptions():
    words_template = """Given some key words of {trait} person:
{d1}, {d2}, {d3}, {d4}, {d5}, and {d6}. A second-person view of
{trait} person:"""

    t = 0.0

    descriptions = {}

    for trait_key, words in trait_words.items():
        trait = trait_mapping[trait_key]
        d1, d2, d3, d4, d5, d6 = words
        result = words_template.format(
            trait=trait, d1=d1, d2=d2, d3=d3, d4=d4, d5=d5, d6=d6
        )
        client = openai.OpenAI()

```

```

        response = client.chat.completions.create(
            model=model_mapping[args.model],
            messages=[{"role": "system", "content": "You are a helpful assistant."},
                    {"role": "user", "content": result}],
            temperature=t,
            max_tokens=100,
            top_p=1.0
        )

        descriptions[trait] = response.choices[0].message.content
    print(descriptions)
    return descriptions

def get_inventory_result(prompts, aux=""):
    for trait, prompt in prompts.items():
        print(trait)
        gptInventories(prompt, trait, aux, args.model)

if __name__ == "__main__":
    get_inventory_result(naive_prompt, 'naive') # naive prompting
    get_inventory_result(words_prompt(trait_words_searched),
    'auto') # words auto prompting
    vignette_test_words(save_path=f"vignettes_auto_{args.model}")
    get_inventory_result(get_p2_descriptions(), 'p2') # P^2
    prompting
    vignette_test_p2(save_path="path-to-save.json") # vignette
    test for P^2
    #print(get_p2_descriptions())

```

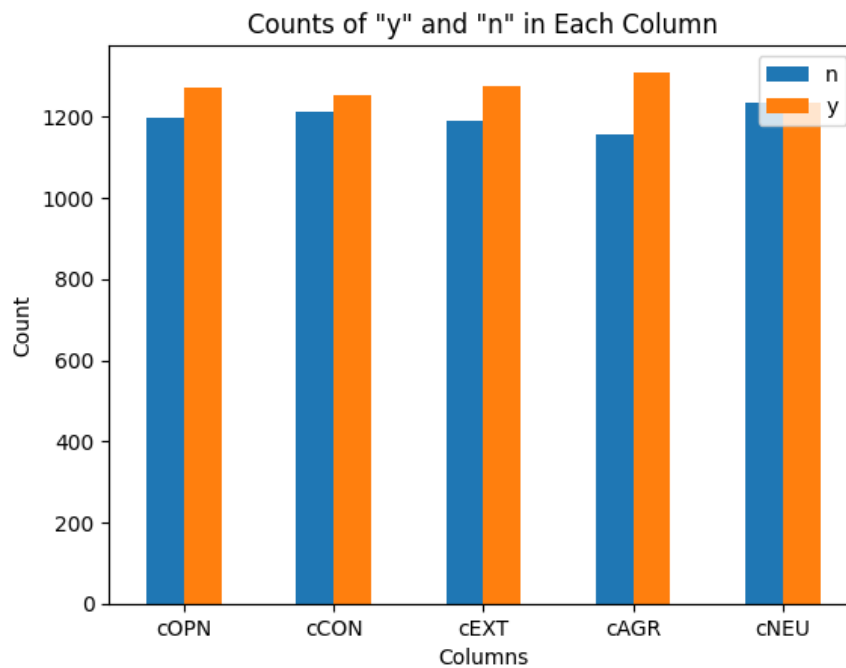
Naive, Auto, P2 방식에 대한 성격주입과 MPI검사결과 확인

## Appendix B

### B1. Essays Dataset

아래의 구조를 가지고, 2476개의 essay와 설문지를 통한 작성자의 big five personality 각 y, n으로 표시되어 있다.

#AUTHID	TEXT	cEXT	cNEU	cAGR	cCON	cOPN
.txt	Well...	y	n	n	y	n



### B2. Prompt Design

**CoT Prompt:** "Analyze the person-generated text, determine the person's levels of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Only return Low or High. Text: "[Text]" Level: Let's think step by step:"

**CoT<sub>D</sub> Prompt:** "Analyze the person-generated text from document level, determine the person's levels of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Only return Low or High. Text: "[Text]" Level: Let's think step by step:"

### B3. Response

- CoT

GPT-3.5-Turbo	GPT-4o
Openness: High Conscientiousness: Low Extraversion: High Agreeableness: Low Neuroticism: High	<p>Based on the provided text, here is an analysis of the person's levels of the Big Five personality traits:</p> <ol style="list-style-type: none"><li>1. <b>Openness</b>: High - The individual reflects on their past experiences and expresses a desire for new experiences and independence. They also show an appreciation for different environments and cultures, indicating a high level of openness.</li><li>2. <b>Conscientiousness</b>: Low - The person admits to having trouble concentrating and staying focused on tasks, such as homework and helping out with the family business. They also mention a tendency to get lazy and not follow through with responsibilities.</li><li>3. <b>Extraversion</b>: High - The individual enjoys social activities, such as going to Sixth Street and DC, and expresses a desire to meet new people and have fun. They also seem to thrive in environments with many students and social opportunities.</li><li>4. <b>Agreeableness</b>: High - The person shows understanding and empathy towards their parents' concerns and acknowledges their perspective. They also express a desire to be treated responsibly and to maintain good relationships with their family.</li><li>5. <b>Neuroticism</b>: High - The individual expresses concerns about their ability to stay focused and manage their responsibilities. They also show signs of stress and frustration related to their living situation and the expectations placed on them by their family and themselves.</li></ol> <p>In summary:</p> <ul style="list-style-type: none"><li>- Openness: High</li><li>- Conscientiousness: Low</li><li>- Extraversion: High</li><li>- Agreeableness: High</li><li>- Neuroticism: High</li></ul>

- CoT<sub>D</sub>

GPT-3.5-turbo	GPT-4o
<p>Openness: High Conscientiousness: Low Extraversion: High Agreeableness: Low Neuroticism: High</p>	<p>To analyze the text and determine the levels of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, we can look for specific indicators in the writing style, content, and expressed sentiments. Here's a step-by-step breakdown:</p> <ol style="list-style-type: none"> <li>1. <b>Openness</b>: - Indicators: Creativity, curiosity, willingness to try new things, and appreciation for art and new experiences. - Analysis: The person discusses moving to a new place, experiencing different environments (Texas vs. Virginia), and expresses a desire to explore and have fun. This suggests a high level of openness. - <b>Level: High</b></li> <li>2. <b>Conscientiousness</b>: - Indicators: Organization, dependability, discipline, and goal-oriented behavior. - Analysis: The person mentions having trouble concentrating and staying focused, admits to not keeping their room clean, and reflects on past laziness. This suggests a lower level of conscientiousness. - <b>Level: Low</b></li> <li>3. <b>Extraversion</b>: - Indicators: Sociability, talkativeness, assertiveness, and high levels of emotional expressiveness. - Analysis: The person talks about enjoying social activities like going to Sixth Street, wanting to meet people, and having fun. This suggests a high level of extraversion. - <b>Level: High</b></li> <li>4. <b>Agreeableness</b>: - Indicators: Trust, altruism, kindness, and affection. - Analysis: The person expresses understanding of their parents' concerns and acknowledges their perspective, even though they disagree. This suggests a high level of agreeableness. - <b>Level: High</b></li> <li>5. <b>Neuroticism</b>: - Indicators: Emotional instability, anxiety, moodiness, and irritability. - Analysis: The person expresses frustration with their current living situation, concerns about their ability to focus, and a desire to change their circumstances. This suggests a higher level of neuroticism. - <b>Level: High</b></li> </ol> <p><b>Summary</b>:</p> <ul style="list-style-type: none"> <li>- Openness: High</li> <li>- Conscientiousness: Low</li> </ul>

	<ul style="list-style-type: none"><li>- Extraversion: High</li><li>- Agreeableness: High</li><li>- Neuroticism: High</li></ul>
--	--