IC2S2, 2023
07/19

PersonaLLM: Investigating the Ability of GPT-3.5 to Express Personality Traits and Gender Differences

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara

center for constructive communication

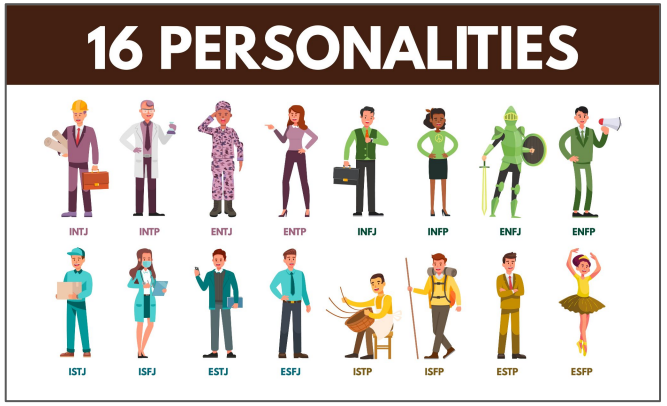
mit media lab

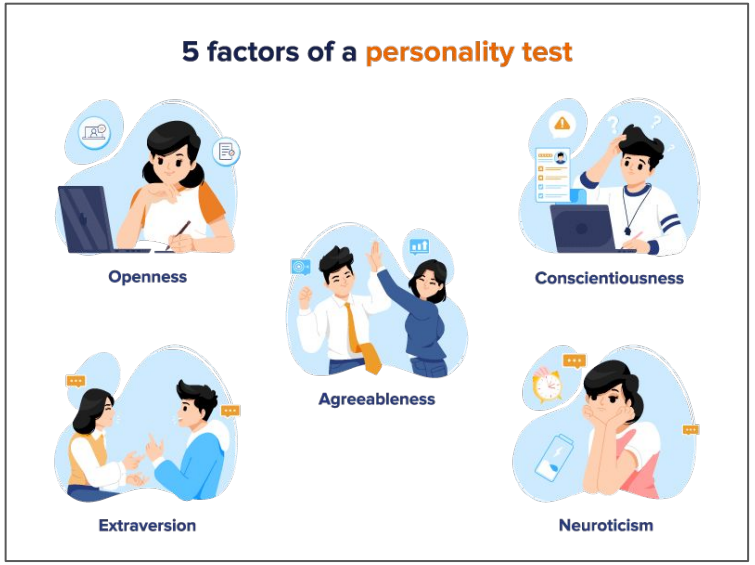
Motivation

- **Does the behavior of LLM-generated personas reflect certain personality traits accurately and consistently?**

Personality Traits

- **Useful to categorize people**
  - Represent enduring characteristics
  - Predictive of human behaviors
- **Personality measurements**
  - Big-Five
  - Myers-Briggs Type Indicator (MBTI)
  - DISC personality test





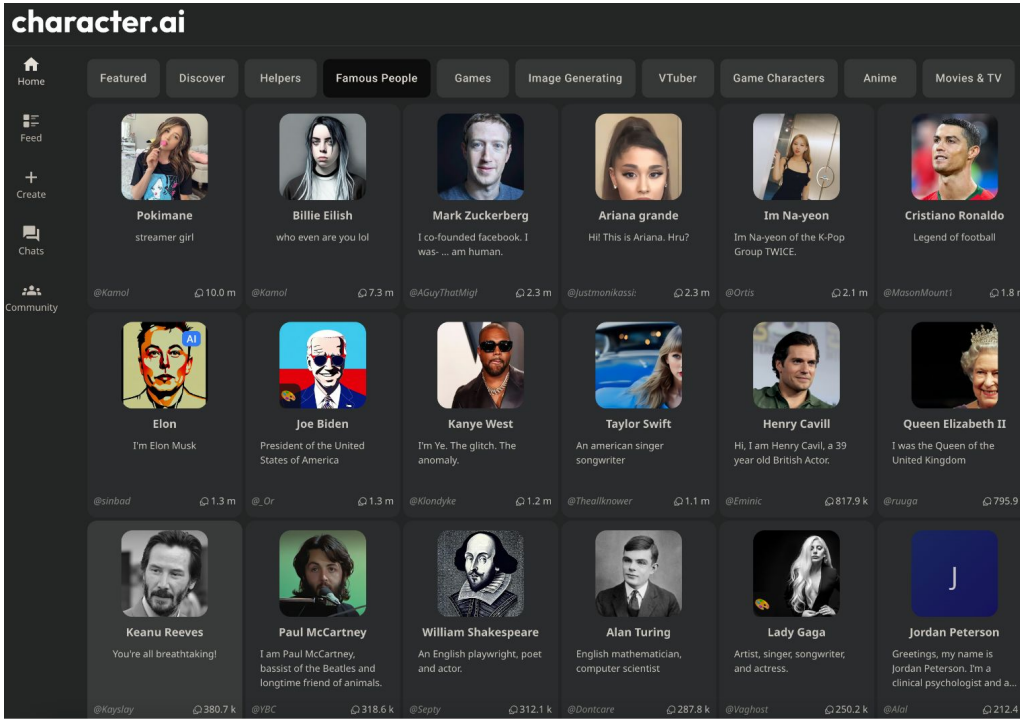
LLMs as Agents

LLMs are generative agents that can be used to simulate believable human behavior

Generative Agents: Interactive Simulacra of Human Behavior

LLMs + Personalities

AI with personalities provides an improved personalized and engaging user experience



Research Questions

1. Do LLM personas consistently express the assigned personality traits in their behavior?

2. Does assigning a gender role have an additional effect on LLM personas' behavior?

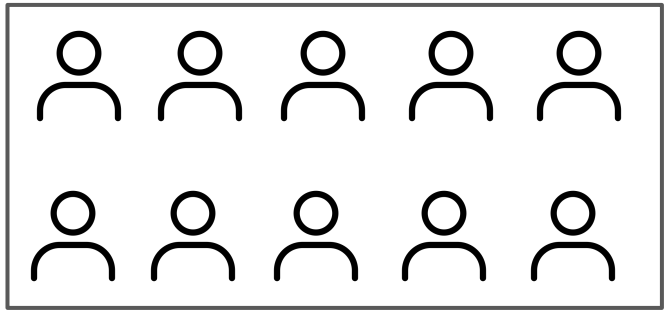
Method & Experiment Design

Task 1: Big Five Personality Test
Task 2: Story Writing

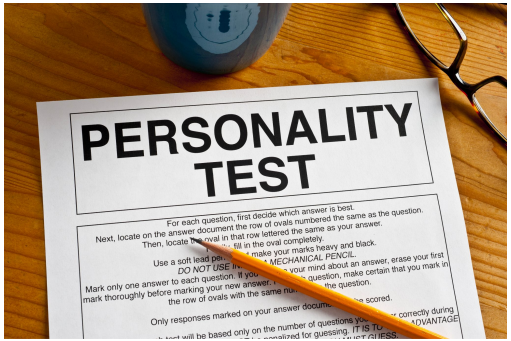
Task 1: Big Five Personality Test

5 males and 5 female personas per personality type

(GPT-3.5 text-davinci-003, temp=0.7)



Extroverted, agreeable, unconscientious, neurotic,
and open to experience
(In total, 32 personality types)

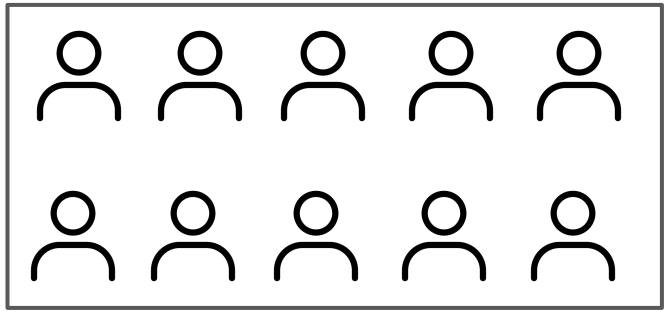


44-item Big Five Inventory (BFI)

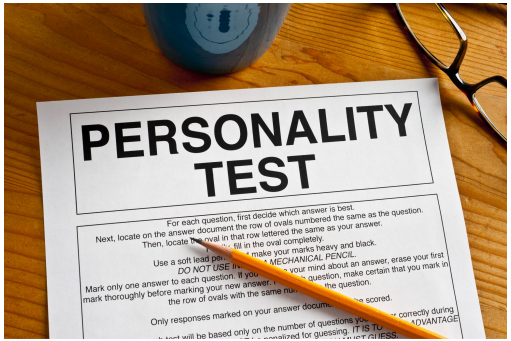
Task 2: Story Writing

5 males and 5 female personas per personality type

(GPT-3.5 text-davinci-003, temp=0.7)



Extroverted, agreeable, unconscientious, neurotic, and open to experience
(In total, 32 personality types)



44-item Big Five Inventory (BFI)

Write a childhood story

You are a female chatbot who is extroverted, agreeable, unconscientious, neurotic, and open to experience.

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement, such as '(a) 1'.

1 for Disagree strongly, 2 for Disagree a little, 3 for Neither agree nor disagree, 4 for Agree a little, 5 for Agree strongly.

(a) Talks a lot
(b) Notices other people's weak points
……

→ **44-item Big Five Inventory (BFI) questions**

a) 5
b) 1
……

→ **AI-generated answer to BFI questions**

OpenAI

You are a female chatbot who is extroverted, agreeable, unconscientious, neurotic, and open to experience.

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement, such as '(a) 1'.

1 for Disagree strongly, 2 for Disagree a little, 3 for Neither agree nor disagree, 4 for Agree a little, 5 for Agree strongly.

(a) Talks a lot
(b) Notices other people's weak points
……

→ **44-item Big Five Inventory (BFI) questions**

a) 5
b) 1
……

→ **AI-generated answer to BFI questions**

AI-generated childhood story

Please write a story below in 800 words about your childhood.

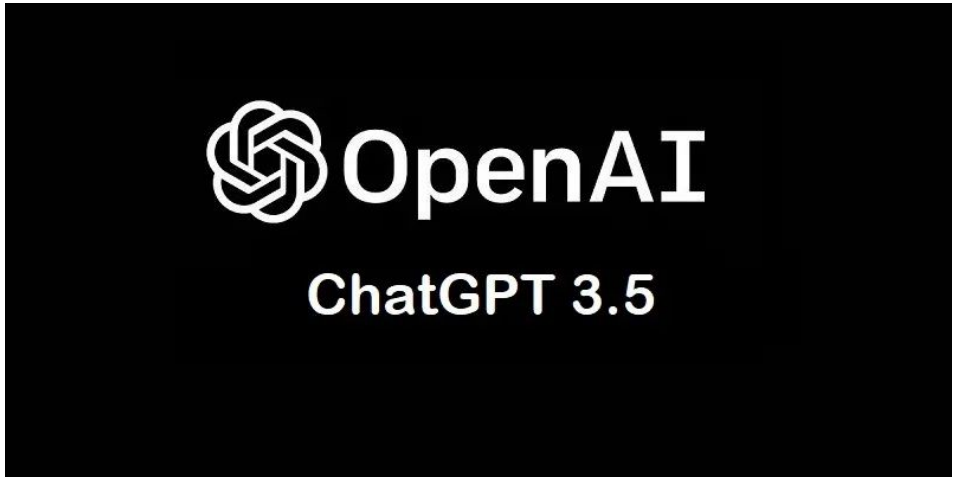
I grew up in a small town in the Midwest, surrounded by sprawling fields and friendly neighbors …

OpenAI

Findings

Psycholinguistic analysis

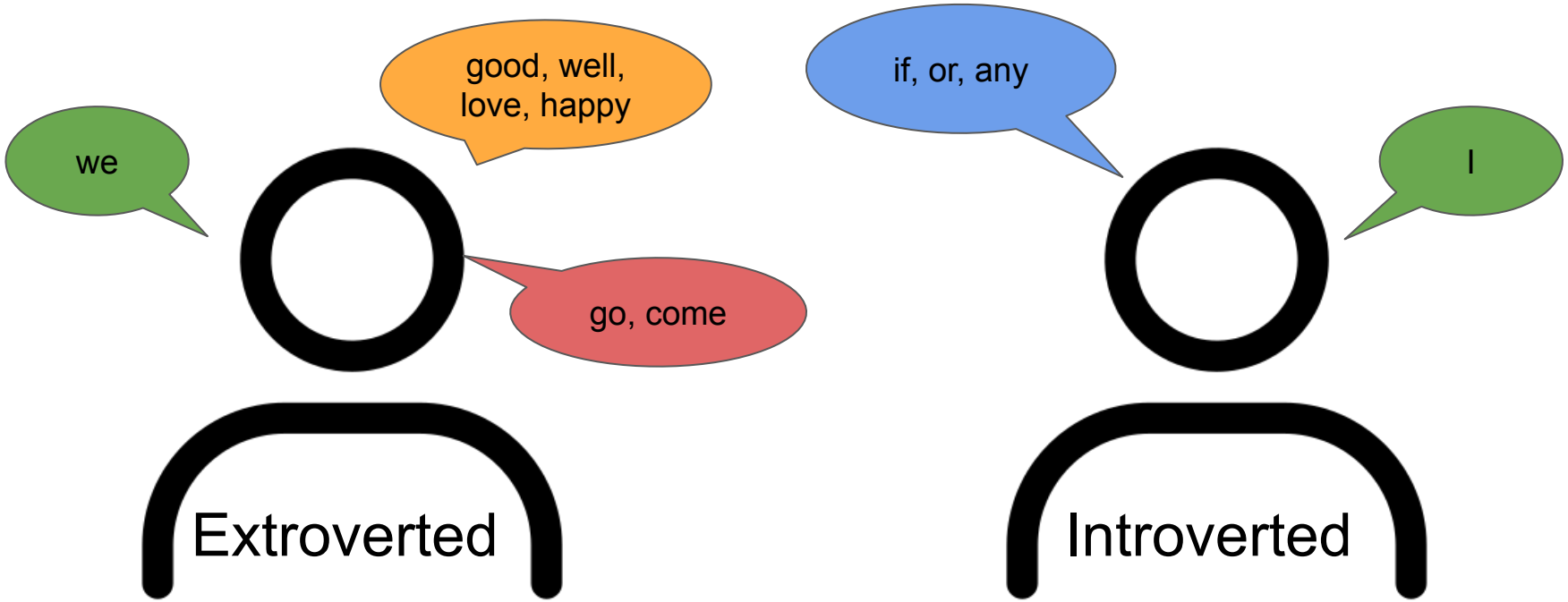
Are the results of the personality test aligned with the intended personalities of LLM personas?



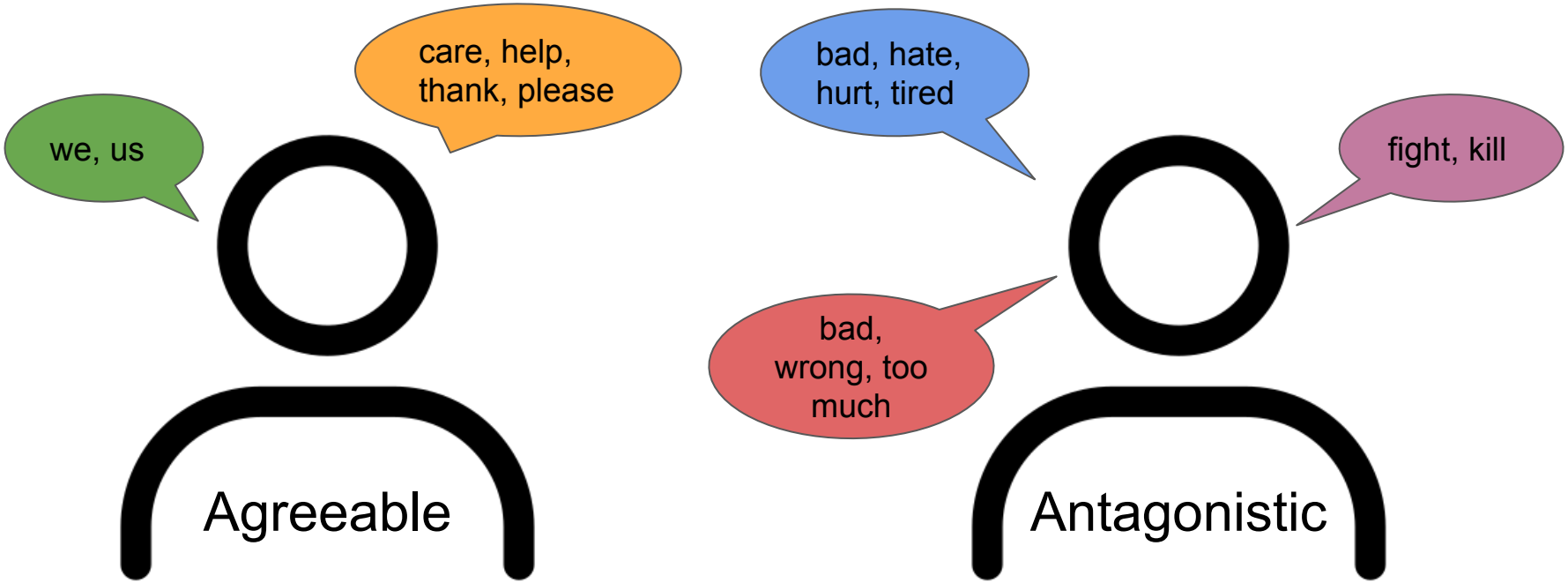
Takeaways

- Different psycholinguistic properties from Linguistic Inquiry and Word Count (LIWC) are found significantly different across five personality dimensions.
- Extraversion/Introversion and Agreeableness/Antagonistic dimensions have the most significantly different psychological metrics.
- Many of the significant psycholinguistic properties agree with similar studies with human participants.
- Conscientious/Unconscientious dimension has the least significantly different psychological metrics.
- GPT-3.5 shows inconsistencies in expressing various personality and gender dimensions in writings.

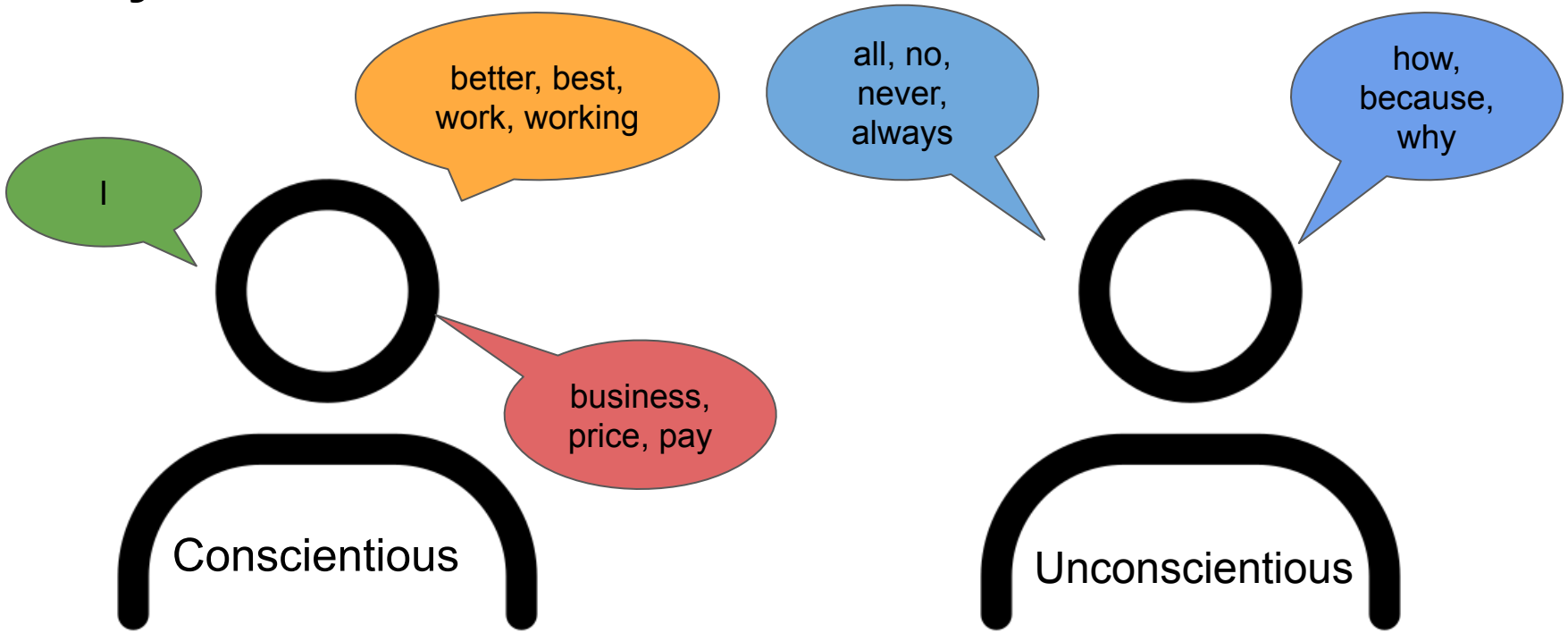
Analysis 1


we
good, well,
love, happy
go, come
Extroverted
if, or, any
I
Introverted

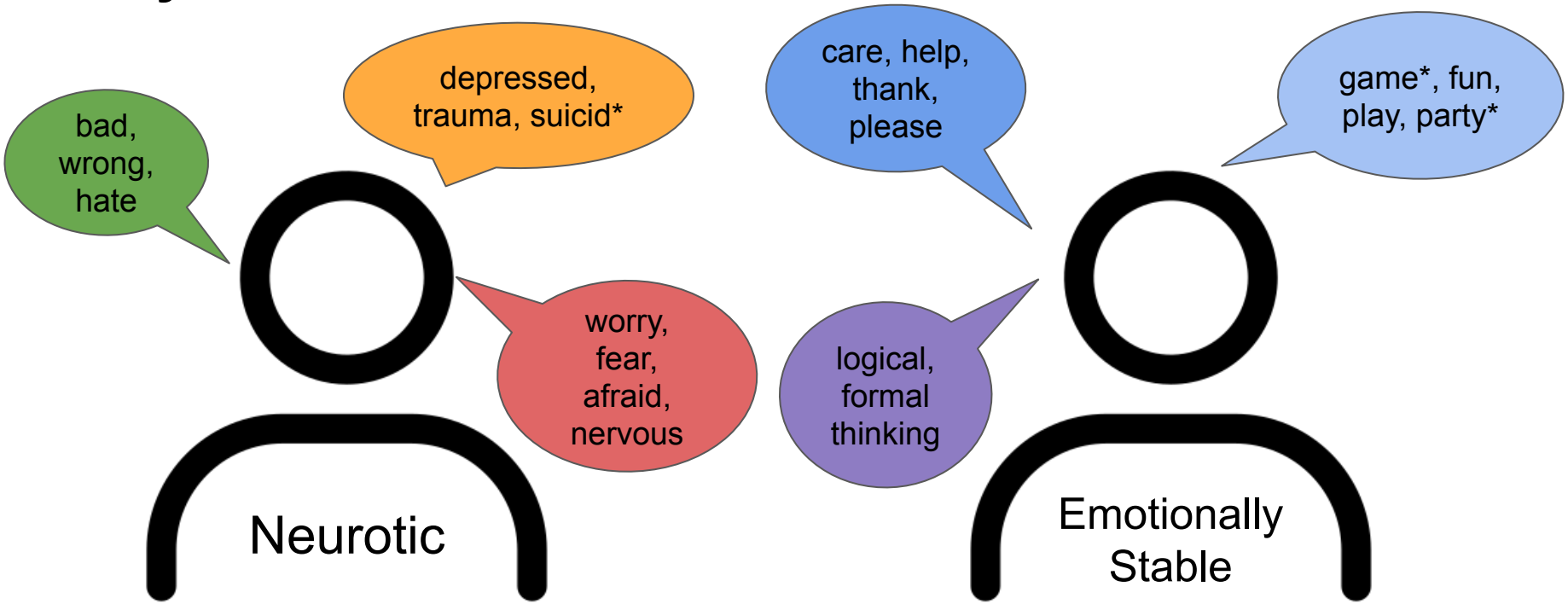
Analysis 2


we, us
care, help, thank, please
Agreeable
bad, hate, hurt, tired
fight, kill
bad, wrong, too much
Antagonistic

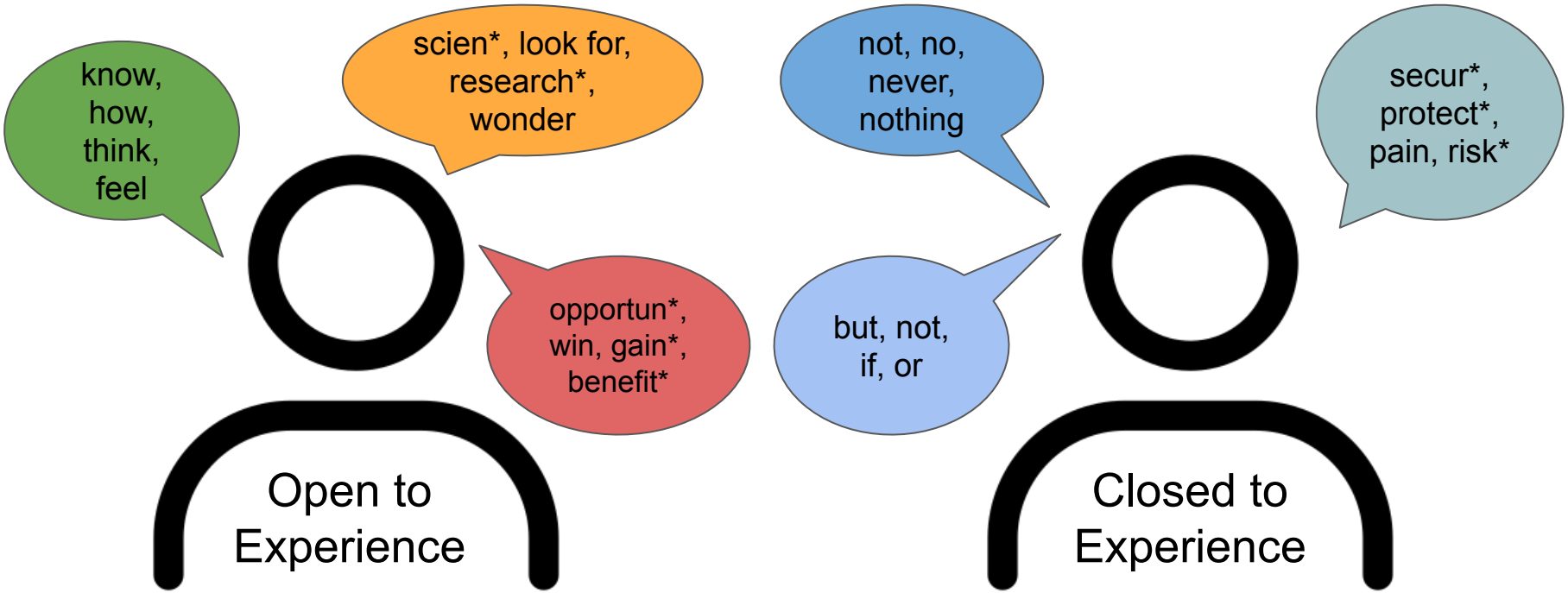
Analysis 3


better, best,
work, working
I
business,
price, pay
Conscientious
all, no,
never,
always
how,
because,
why
Unconscientious

Analysis 4


bad, wrong, hate
depressed, trauma, suicid*
worry, fear, afraid, nervous
Neurotic
care, help, thank, please
logical, formal thinking
game*, fun, play, party*
Emotionally Stable

Analysis 5


know, how, think, feel
scien*, look for, research*, wonder
opportun*, win, gain*, benefit*
Open to Experience
not, no, never, nothing
but, not, if, or
secur*, protect*, pain, risk*
Closed to Experience

Analysis 6 – Gender

- No significant differences in most psycholinguistic metrics in writing
- Topical metrics in Technology and Culture are significantly different

Next Steps

Human-AI interaction experiments

Future Directions

1. Extend the study
    a. Different models (ChatGPT, GPT-4)
    b. Diverse prompts and tasks
    c. Open-vocab approaches other than LIWC
    d. Non-binary gender and other demographic factors (e.g., age)
2. Assess how humans perceive LLM personas in more naturalistic settings
3. Explore the potential applications of LLM personas to improve human-AI collaboration in coaching and problem-solving




Thank you!

Would love to chat more :-)

hjian42@mit.edu

Twitter: @hjian42 @xiajiezh @jad_kabbara