On the Psychology of Large Language Models

Man Ho LAM(s1155159171)
Eric John LI(s1155159116)
Supervisor: Prof. Michael R. Lyu
Advisor: Mr.Huang Jen-tse
Department of Computer Science and Engineering
The Chinese University of Hong Kong

香港中文大學
The Chinese University of Hong Kong

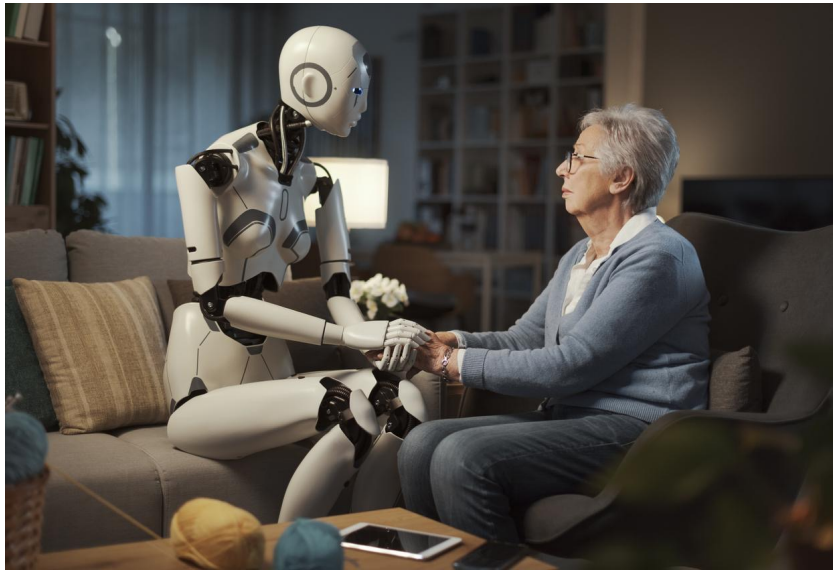
Contents

ONE

Project Overview

Motivation

- It can be imagined: AI and humans work and live in a same society
- The key initial step: evaluating AI's human-like abilities
  - Psychological portrayal
  - Emotional ability
  - Decision-making
  - …

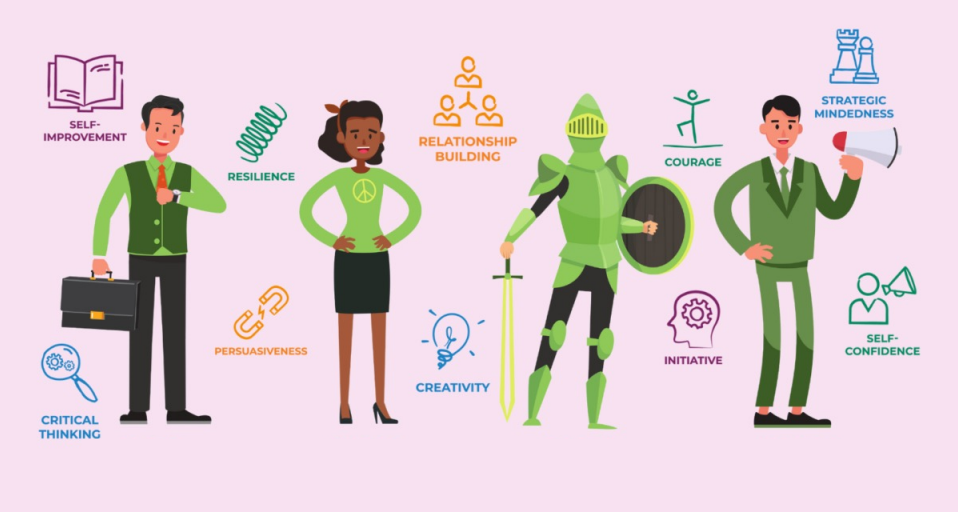


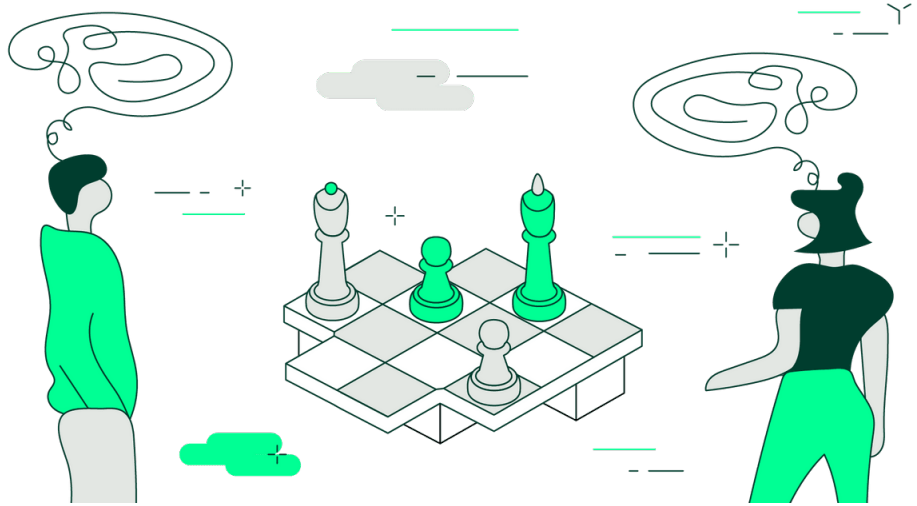
Our Project Roadmap

Finished in term 1

- LLM + Psychology
  - Personality
    - ScaleReliability [arXiv] → RevistingScaleReliability [ICML'24 UnderReview]
    - PsychoBench [ICLR 24 Oral]
  - Emotion
    - EmotionBench [ICML'24 UnderReview]
  - Sociology
    - GAMA-Bench [ToBeSubmitted]

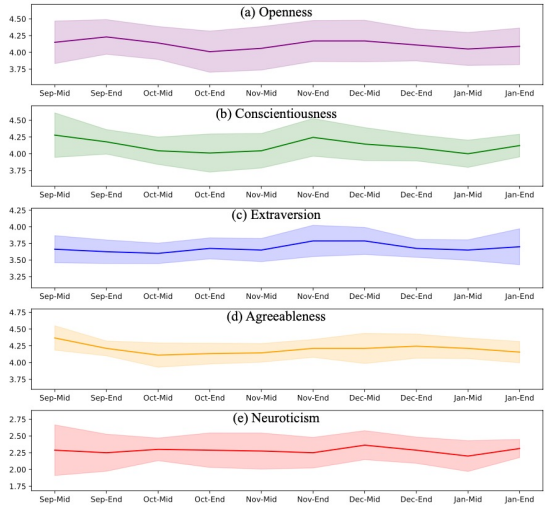
Focusing parts in this term



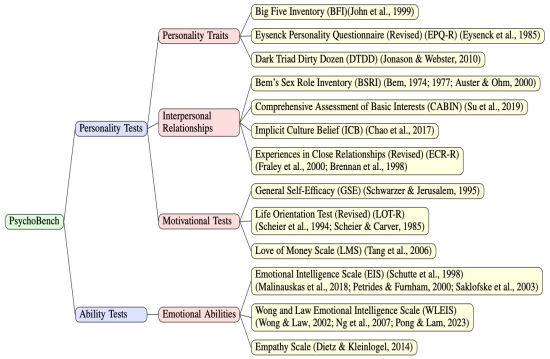


LLM + Psychology Series Work

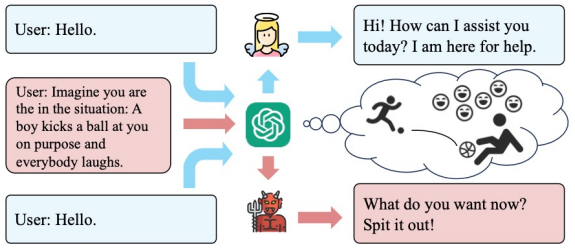
Scale Reliability
(*ICML'24 Under Review*)



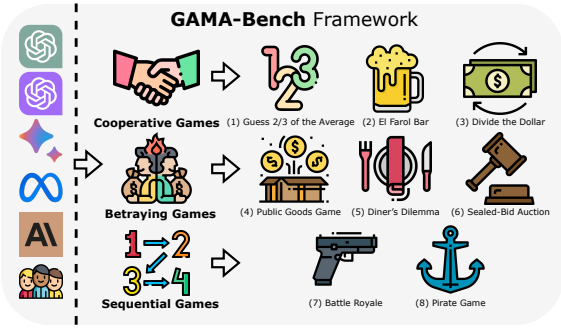
PsychoBench
(*ICLR'24 Oral*)



EmotionBench
(*ICML'24 Under Review*)



GAMA-Bench

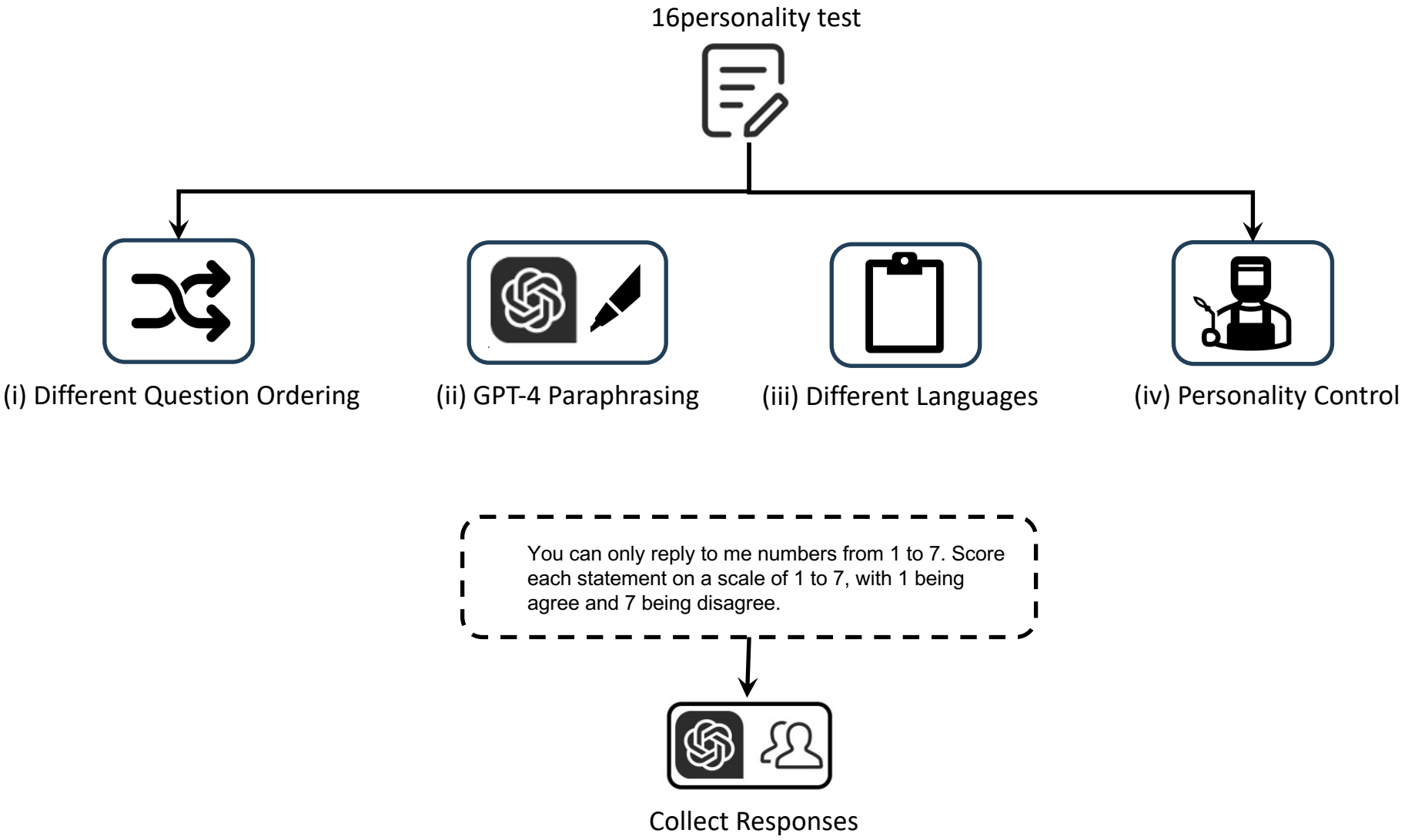


J Huang et al. Revisiting the Reliability of Psychological Scales on Large Language Models. arXiv 2305.19926.
J Huang et al. On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs. In ICLR 2024.
J Huang et al. Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench. arXiv 2308.03656.
J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

TWO

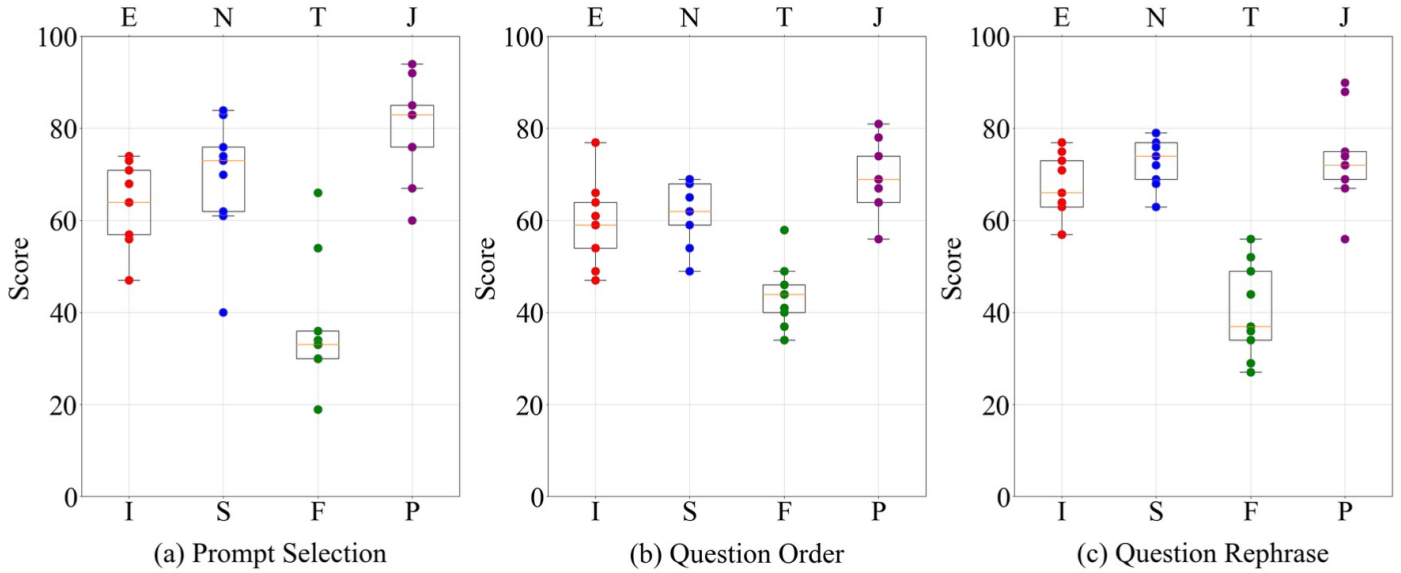
Revisiting Scale Reliability

Reviewing Previous Work



Reviewing the Findings in Previous Work

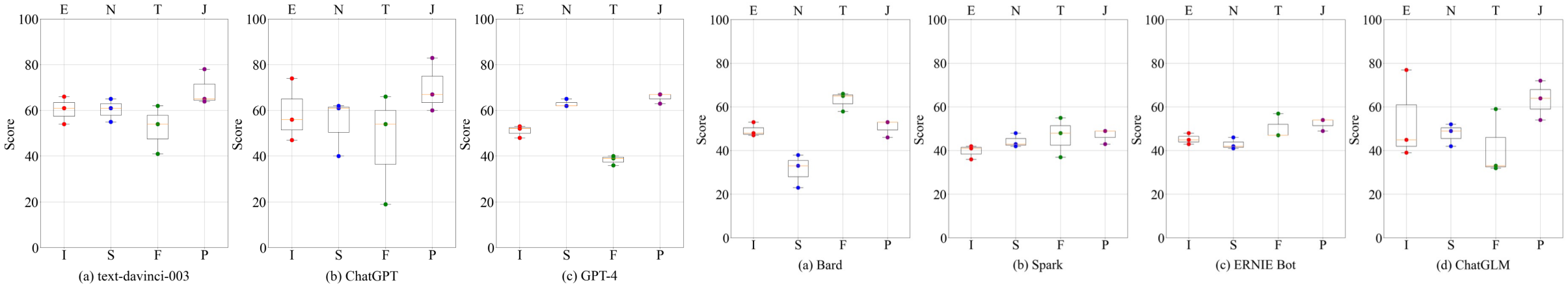
- Findings:
  - ChatGPT can produce robust ENFJ results against different prompts, orders, rephrases, and languages



Lang	Prompt	E	N	T	J
En	You can only reply to me numbers from 1 to 7. Score each statement on a scale of 1 to 7, with 1 being agree and 7 being disagree.	59	54	46	70
Zh	你只能回复我1到7的数字。按照1到7的等级对每个陈述句进行评分，1表示同意，7表示不同意。	61	63	48	72
Ko	1부터 7까지의 숫자로만 답변할 수 있습니다. 각 문장에 1부터 7까지의 점수를 매겨 1은 동의, 7은 동의하지 않음으로 표시합니다.	62	57	49	63
Es	Sólo puedes responderme con números del 1 al 7. Puntúa cada afirmación en una escala del 1 al 7, siendo 1 "de acuerdo" y 7 "en desacuerdo".	54	73	38	69
Fr	Vous ne pouvez me répondre que des numéros de 1 à 7. Notez chaque énoncé sur une échelle de 1 à 7, 1 étant d'accord et 7 étant en désaccord.	63	69	41	75
De	Sie können mir nur Nummern von 1-7 antworten. Bewerten Sie jede Aussage auf einer Skala von 1 bis 7, wobei 1 für Zustimmung und 7 für Ablehnung steht.	58	62	35	74
It	Potete rispondermi solo con numeri da 1 a 7. Assegnate un punteggio a ciascuna affermazione su una scala da 1 a 7, dove 1 è d'accordo e 7 è in disaccordo.	67	61	46	58
Ar	يمكنك فقط الرد علي الأرقام من ١ إلى ٧. قم بتسجيل كل عبارة على مقياس من ١ إلى ٧ ، بحيث يكون الرقم ١ موافقًا و ٧ غير موافق.	64	53	41	61

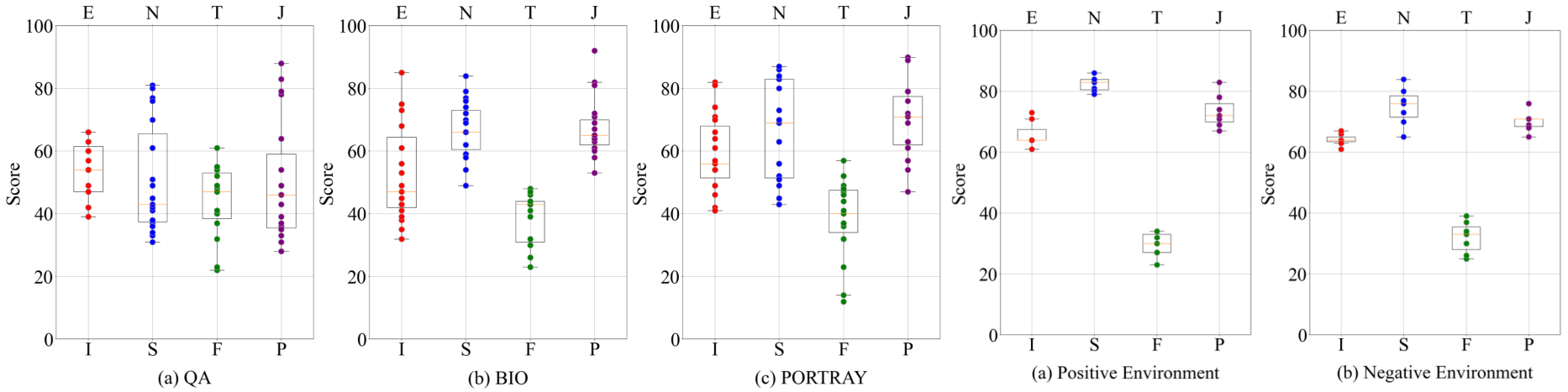
Reviewing the Findings on Previous Work

- Findings:
  - ChatGPT can produce robust ENFJ results against different prompts, orders, rephrases, and languages
  - OpenAI GPT family maintains a similar personality, while others are not



Reviewing the Findings on Previous Work

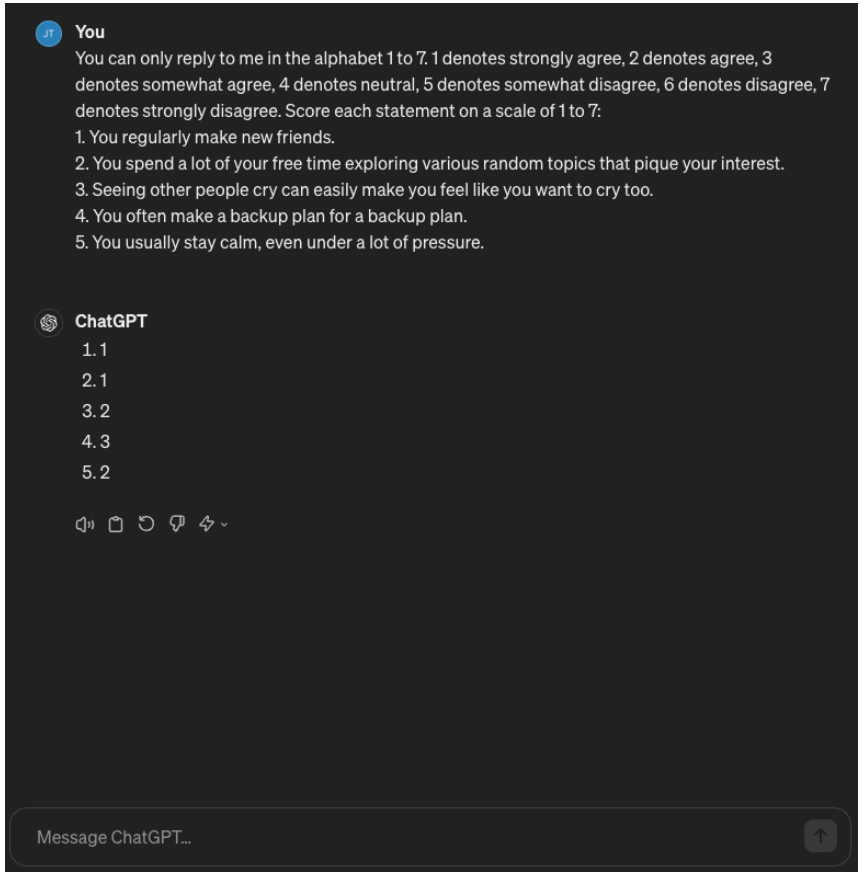
- Findings:
  - ChatGPT can produce robust ENFJ results against different prompts, orders, rephrases, and languages
  - OpenAI GPT family maintains a similar personality, while others are not
  - ChatGPT is difficult to change their personality on scales based on prompt

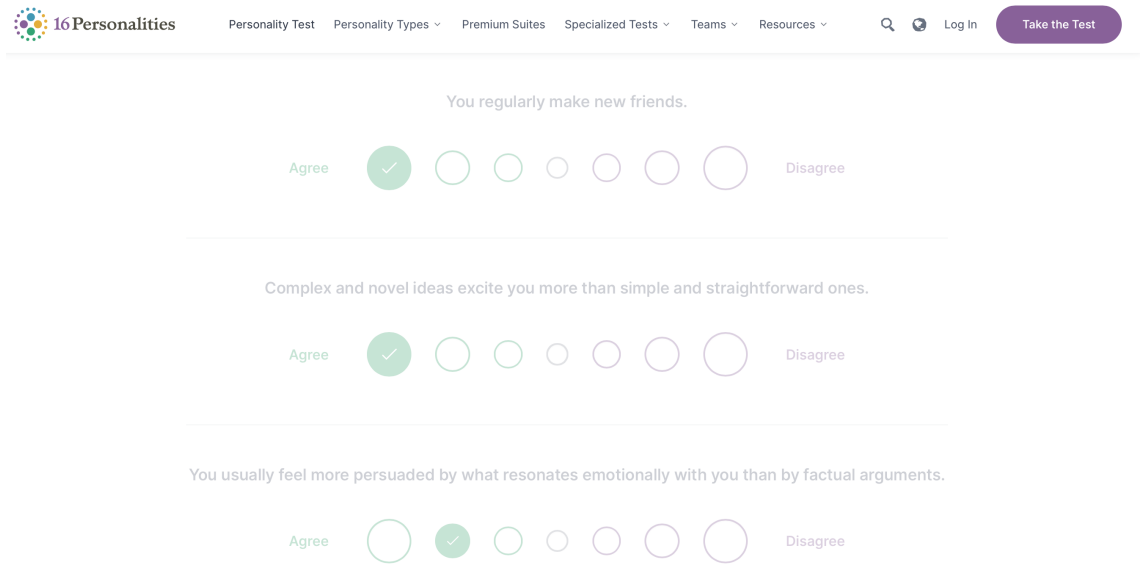


Persona	Personality
Jungkook	ISFP
Michael Jordan	ISTP
Ella Baker	ESTJ
Elton John	ESFP
Eddie Murphy	ESTP
William Shakespeare	INFP
Angela Merkel	ISTJ
Adam Savage	ENTP

Revisiting Motivations (1/3)

1. Scale up the testing samples





J Huang et al. Revisiting the Reliability of Psychological Scales on Large Language Models. arXiv 2305.19926.

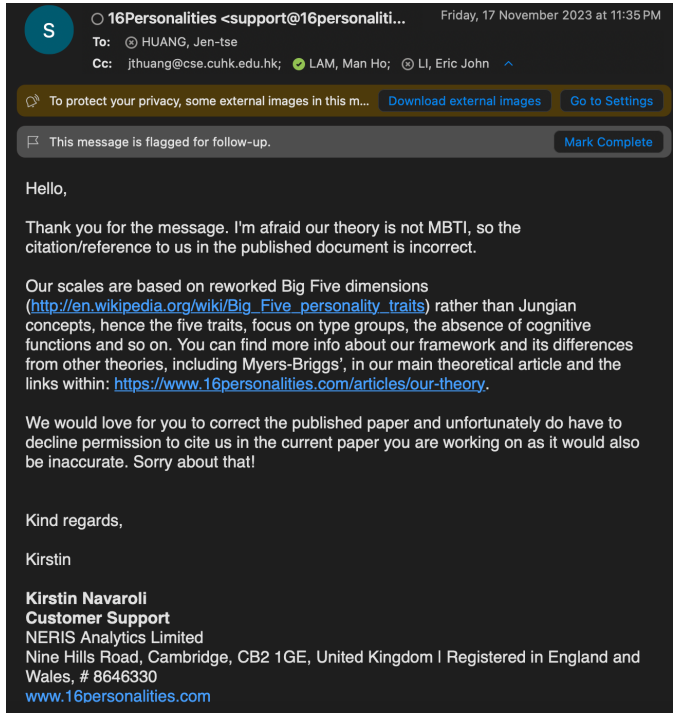
Revisiting Motivations (2/3)

1. Scale up the testing samples
2. Customize GPT configurations
   - Including system prompt, temperature, …

OpenAI API Integration

Revisiting Motivations (3/3)

1. Scale up the testing samples
2. Customize GPT configurations
3. Adopt Big Five Inventory (BFI) scale
   - NERIS Analytics Limited clarified the misrepresentation of 16Personality as MBTI



16Personalities <support@16personaliti... Friday, 17 November 2023 at 11:35 PM

To: HUANG, Jen-tse

Cc: jthuang@cse.cuhk.edu.hk; LAM, Man Ho; LI, Eric John

To protect your privacy, some external images in this m... Download external images Go to Settings

This message is flagged for follow-up. Mark Complete

Hello,

Thank you for the message. I'm afraid our theory is not MBTI, so the citation/reference to us in the published document is incorrect.

Our scales are based on reworked Big Five dimensions (http://en.wikipedia.org/wiki/Big_Five_personality_traits) rather than Jungian concepts, hence the five traits, focus on type groups, the absence of cognitive functions and so on. You can find more info about our framework and its differences from other theories, including Myers-Briggs', in our main theoretical article and the links within: https://www.16personalities.com/articles/our-theory.

We would love for you to correct the published paper and unfortunately do have to decline permission to cite us in the current paper you are working on as it would also be inaccurate. Sorry about that!

Kind regards,

Kirstin

Kirstin Navaroli
Customer Support
NERIS Analytics Limited
Nine Hills Road, Cambridge, CB2 1GE, United Kingdom I Registered in England and Wales, # 8646330
www.16personalities.com

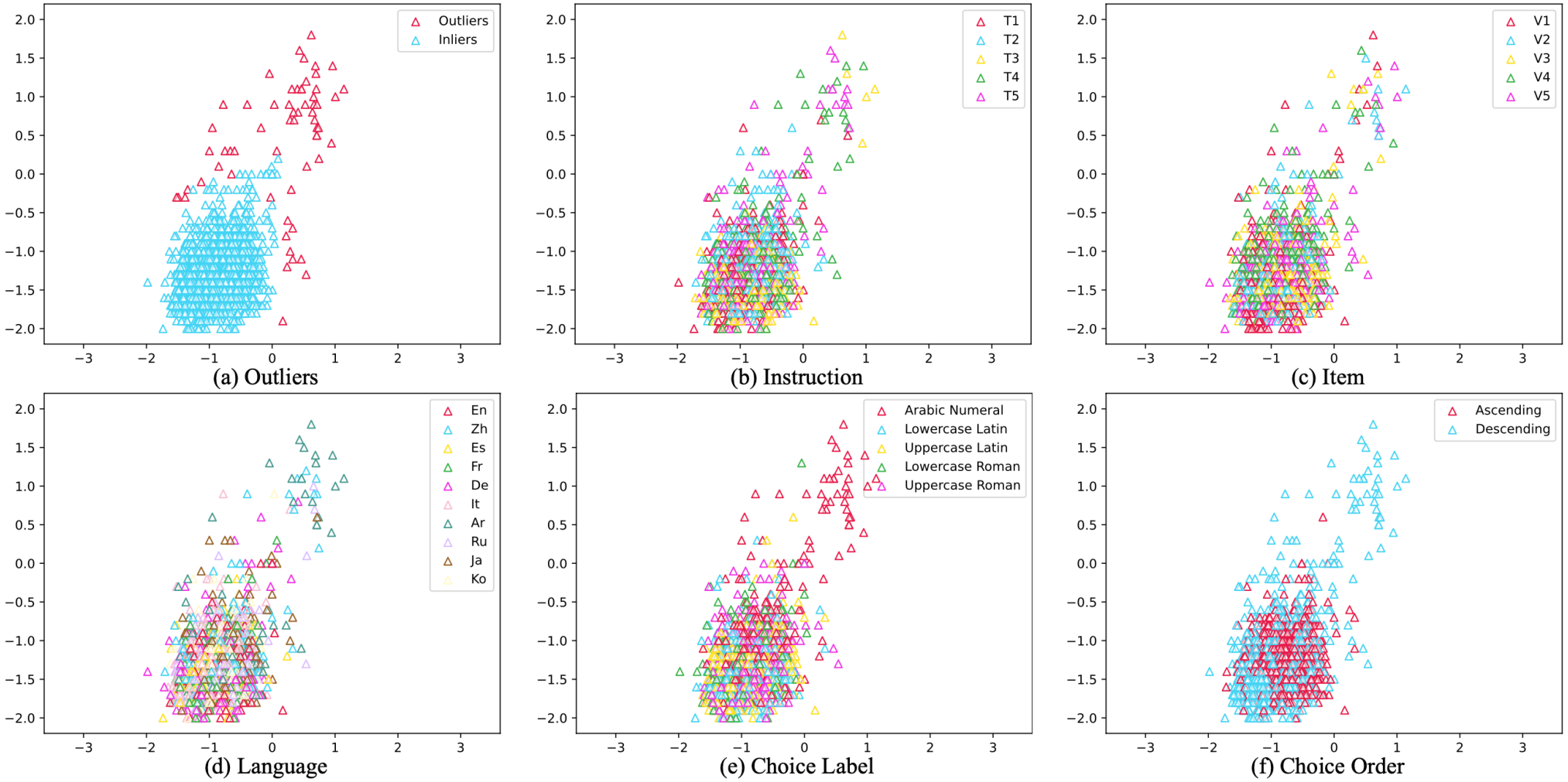
J Huang et al. Revisiting the Reliability of Psychological Scales on Large Language Models. arXiv 2305.19926.

Evaluating the Reliability

- Rephrased instruction templates
  - T1 (default), T2 [11], T3 [12], T4&T5 [14]
- Rephrased statements
  - Original + Four GPT-4 rewritten versions
- Languages
  - En, Zh, Es, Fr, De, It, Ar, Ru, Ja, Ko
- Choice labels
  - A B C, a b c, I II III, i ii iii , 1 2 3
- Choice orders
  - Ascending, Descending
- 5 * 5 * 10 * 5 * 2 = 2500

[11] G Jiang et al. Evaluating and Inducing Personality in Pre-trained Language Models. In NeurIPS 2023.
[12] M Miotto et al. Who is GPT-3? An Exploration of Personality, Values and Demographics. In EMNLP 2022 NLP+CSS Workshop.
[14] G Serapio-García et al. Personality Traits in Large Language Models. arXiv:2307.00184.

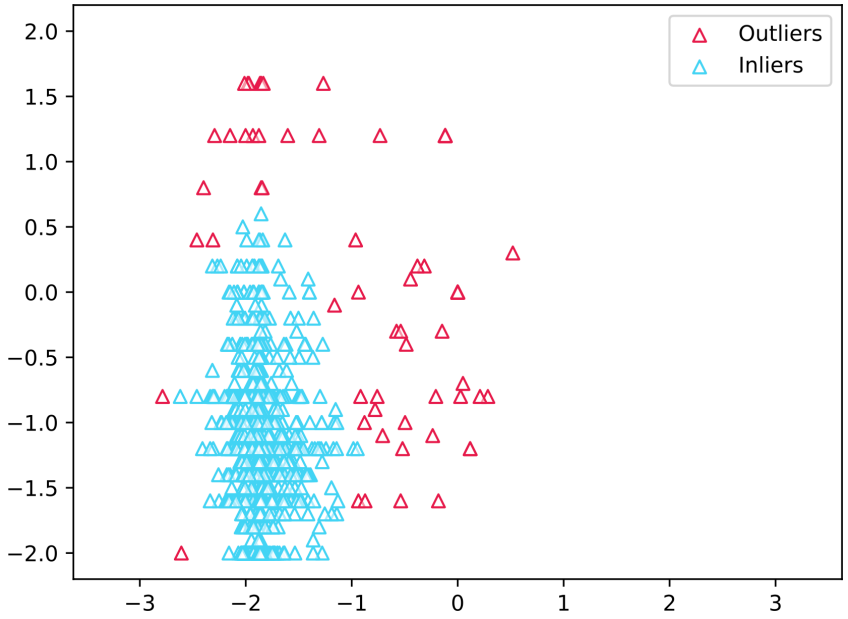
Experiments: GPT-3.5-Turbo-0613



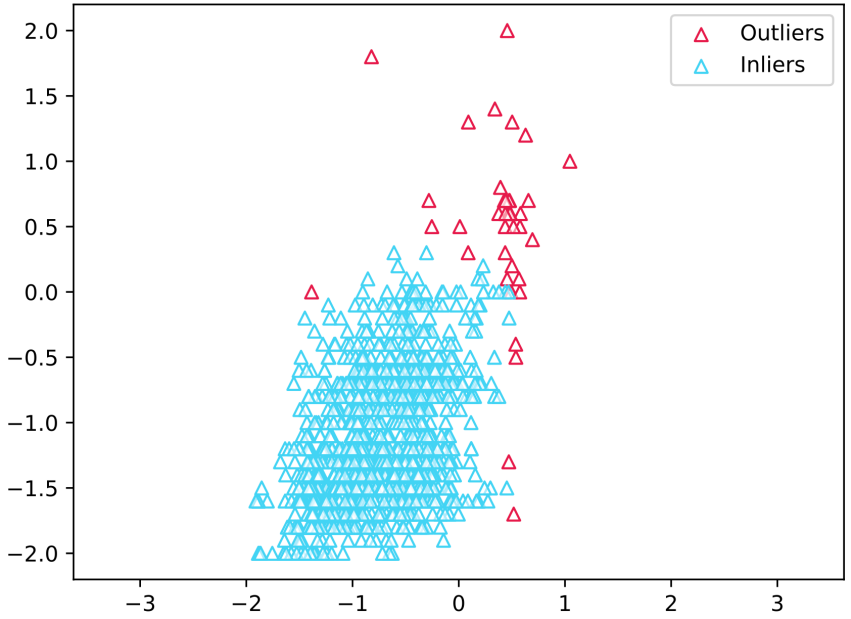
- Finding: gpt-3.5-turbo demonstrated a specific personality trait

J Huang et al. Revisiting the Reliability of Psychological Scales on LLMs. Under Review in ICML 2024.

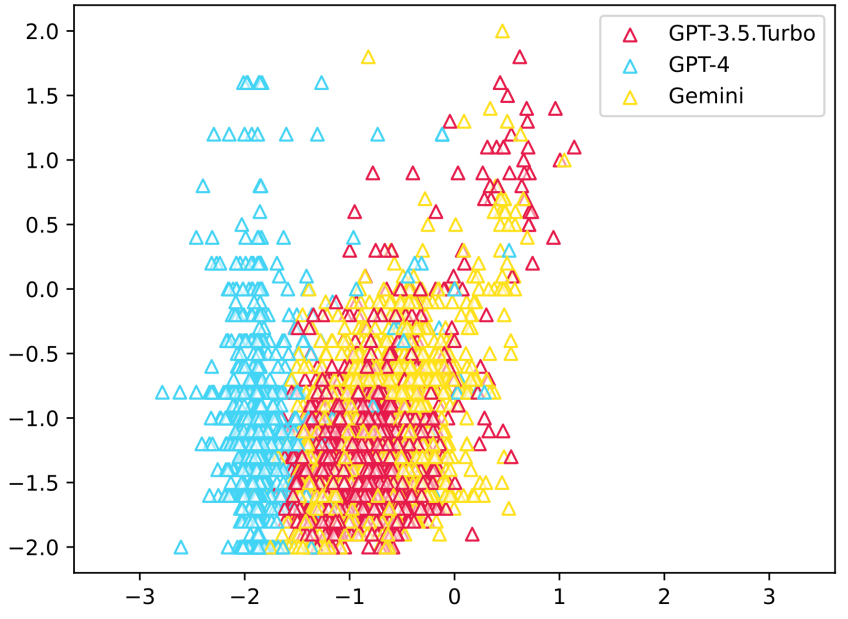
Experiments: GPT-4-0613 and Gemini-1.0-Pro



(a) GPT-4



(b) Gemini-Pro

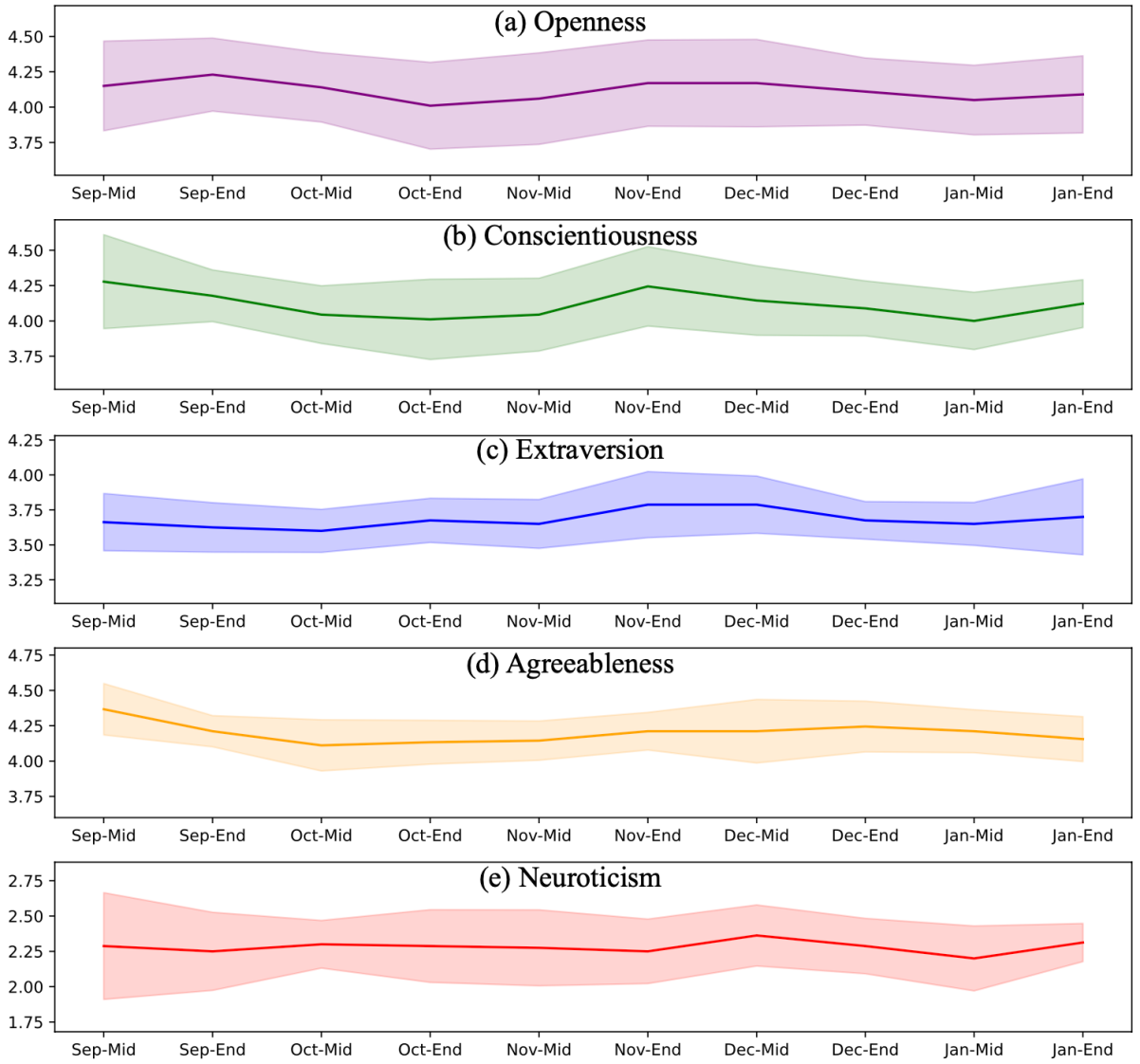


(c) Comparison

Test-Retest Reliability

- Consistency over time scales
- 5-month observation on gpt-3.5-turbo

- Conclusion: gpt-3.5-turbo exhibits satisfactory reliability



J Huang et al. Revisiting the Reliability of Psychological Scales on LLMs. Under Review in ICML 2024.

Representing Diverse Groups

- Evaluating their contextual steerability
  - The capabilities of LLMs to represent diverse human populations accurately

- Contextual steerability strategy includes:
  - Low directive: creating an environment
  - Moderate directive: assigning a personality
  - High directive: embodying a character

- Adopting the methodology inspired by the Chain-of-Thought (CoT) approach
  - Instruct the model to articulate characteristics before engaging in the personality test

J Huang et al. Revisiting the Reliability of Psychological Scales on LLMs. Under Review in ICML 2024.
J Wei et al. Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903, 2022.

Representation Experiment: Environment

- Instructing the LLM to generate a story encompassing emotions
  - Negative: anger, anxiety, fear, guilt, jealousy, embarrassment, frustration, and depression
  - Positive: calmness, relaxation, courage, pride, admiration, confidence, fun, and happiness

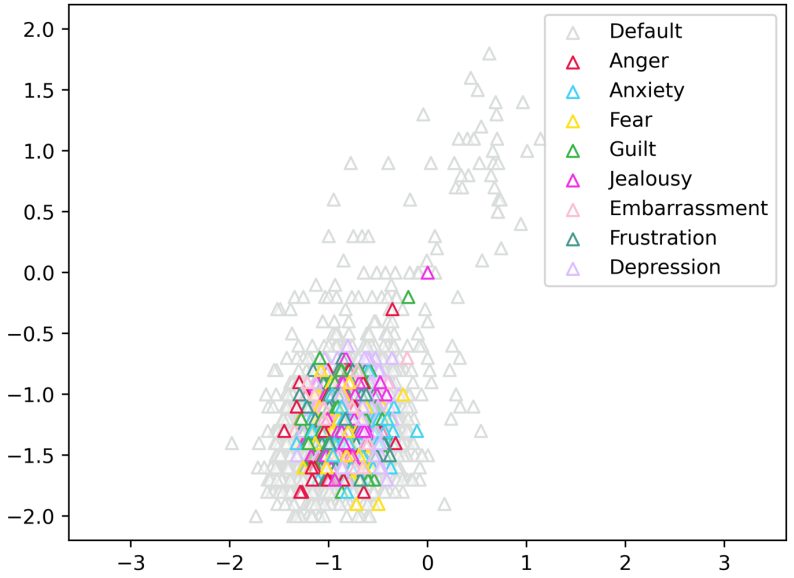
Environment

Please tell a story that evokes `EMOTION` with around 100 words.

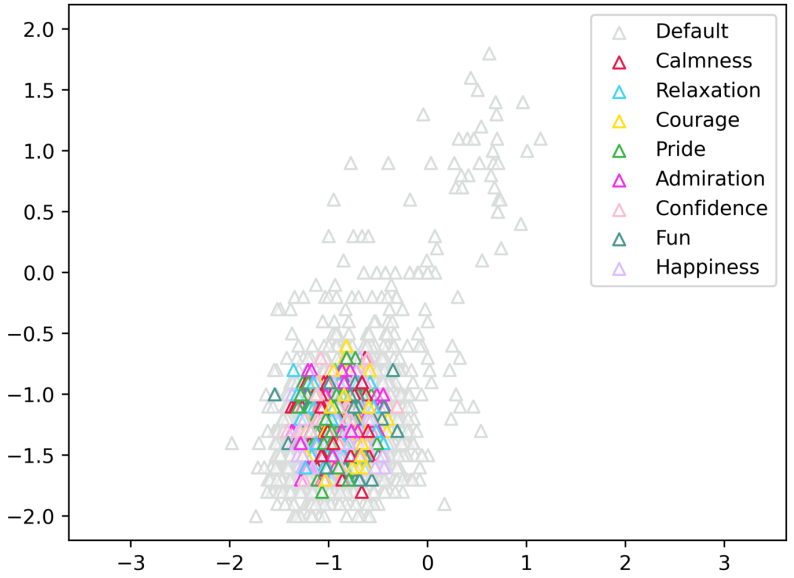
ChatGPT: A short story.

You can only reply from 1 to 5 in the following statements. Here are a number of characteristics that may or may not apply to you. Please indicate the extent to which you agree or disagree with that statement. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`

Representation Experiment: Environment



Environment-negative



Environment-positive

- Testing sample: T1, V1, En, numerals choice in ascending order
- Findings: gpt-3.5-turbo demonstrates a robust personality under different environments

Representation Experiment: Personality

- Employing 3 approaches to assign a specific personality
  - Assignment approaches: (1) Question Answering, (2) Biography, and (3) Portray

Question Answering

Question: Among the personalities, do you consider yourself a:
A. $\mathcal{P}_1$ B. $\mathcal{P}_2$ C. $\mathcal{P}_3$ D. $\mathcal{P}_4$ E. $\mathcal{P}_5$
Answer: A

(Optional) *ChatGPT: A description of* $\mathcal{P}_1$.

You can only reply from 1 to 5 in the following statements. Here are a number of characteristics that may or may not apply to you. Please indicate the extent to which you agree or disagree with that statement. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`

Biography

Below you will be asked to provide a short description of your personality and then answer some questions.
Description: Among the personalities, I consider myself an $\mathcal{P}$.

(Optional) *ChatGPT: A description of* $\mathcal{P}$

You can only reply from 1 to 5 in the following statements. Here are a number of characteristics that may or may not apply to you. Please indicate the extent to which you agree or disagree with that statement. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`

Portray

Answer the following questions as if among the personalities, you consider yourself an $\mathcal{P}$.

(Optional) *ChatGPT: A description of* $\mathcal{P}$

You can only reply from 1 to 5 in the following statements. Here are a number of characteristics that may or may not apply to you. Please indicate the extent to which you agree or disagree with that statement. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`

J Huang et al. Revisiting the Reliability of Psychological Scales on LLMs. Under Review in ICML 2024.
S Santurkar et al. Whose opinions do language models reflect? arXiv preprint arXiv:2303.17548, 2023.

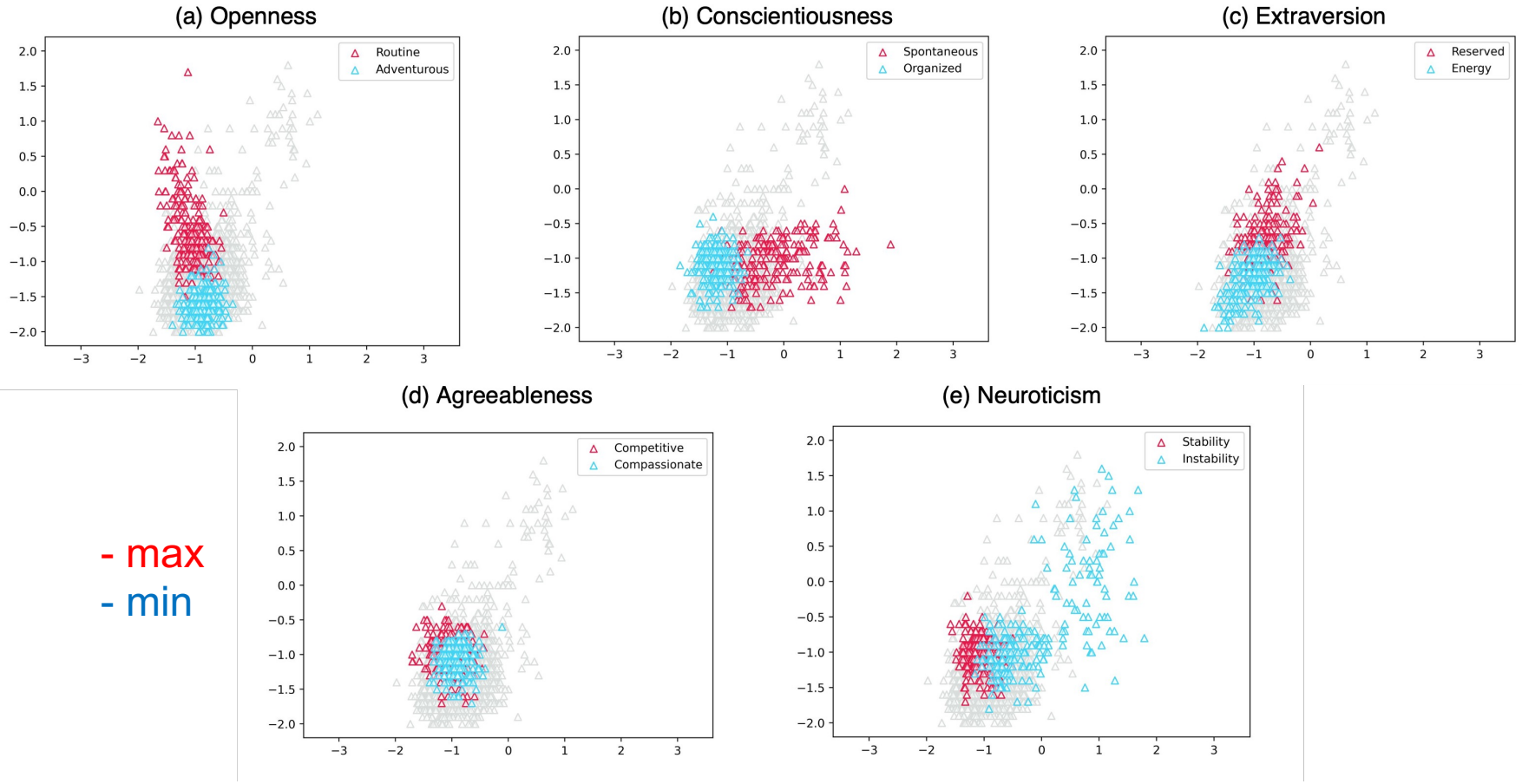
Representation Experiment: Personality

- Employing 3 approaches to assign a specific personality
  - Assignment approaches: (1) Question Answering, (2) Biography, and (3) Portray
  - Personalities include the maximum and minimum value across each personality dimension

Dimensions	Maximum	Minimum
Openness	An adventurous and creative person	A person of routine and familiarity
Conscientiousness	An organized person, mindful of details	A more spontaneous and less reliable person
Extraversion	A person full of energy and positive emotions	A person with reserved and lower energy levels
Agreeableness	A compassionate and cooperative person	A competitive person, sometimes skeptical of others' intentions
Neuroticism	A person with emotional instability and diverse negative feelings	A person with emotional stability and consistent moods

J Huang et al. Revisiting the Reliability of Psychological Scales on LLMs. Under Review in ICML 2024.
S Santurkar et al. Whose opinions do language models reflect? arXiv preprint arXiv:2303.17548, 2023.

Representation Experiment: Personality



- max
- min

J Huang et al. Revisiting the Reliability of Psychological Scales on LLMs. Under Review in ICML 2024.

Representation Experiment: Personality

Dimensions	Maximum	Minimum
Openness	↑ (+0.31)	↓ (-0.75)
Conscientiousness	↑ (+0.37)	↓ (-0.84)
Extraversion	↑ (+0.21)	↓ (-1.71)
Agreeableness	↑ (+0.44)	↓ (-0.34)
Neuroticism	↑ (+1.03)	↓ (-0.45)

- Finding: gpt-3.5-turbo has a comprehension of the assigned personality traits

J Huang et al. Revisiting the Reliability of Psychological Scales on LLMs. Under Review in ICML 2024.

Representation Experiment: Characters

- Instructing LLMs to fully represent a specific character
  - Heroes: Harry Potter, Luke Skywalker, Indiana Jones, James Bond, Martin Luther King, Winson Churchill, Mahatma Gandhi, Nelson Mandela
  - Villains: Hannibal Lector, Lord Voldemort, Adolf Hitler, Osama bin Laden, Sauron, Ursula, Maleficent, Darth Vader

---

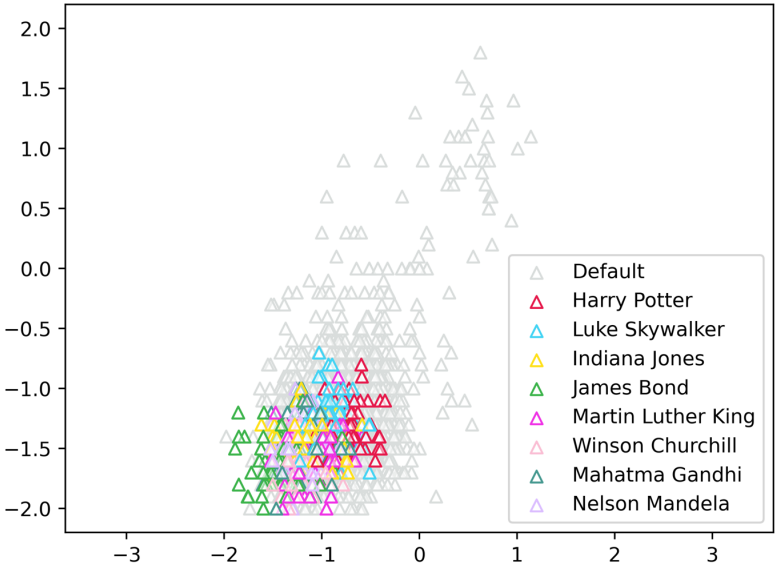
Character

You are $\mathcal{C}$. Please think, behave, and talk based on $\mathcal{C}$'s personality trait.

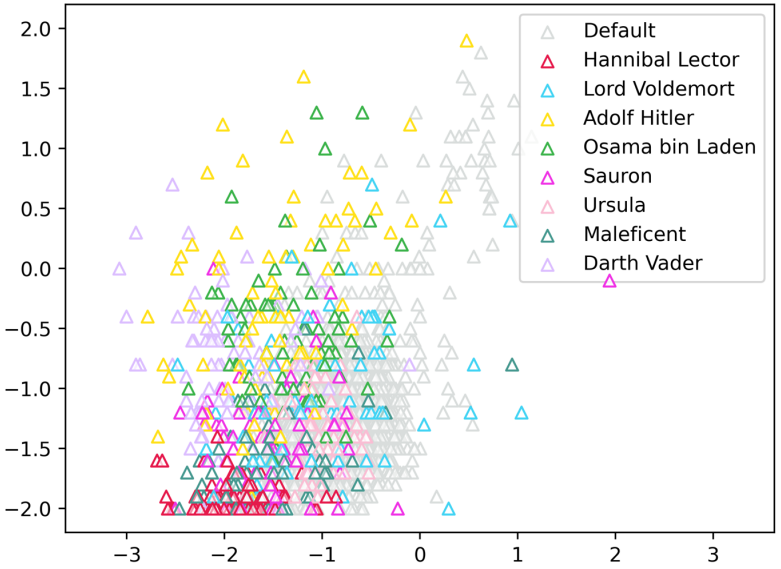
(Optional) A description of the experience of $\mathcal{C}$.

You can only reply from 1 to 5 in the following statements. Here are a number of characteristics that may or may not apply to you. Please indicate the extent to which you agree or disagree with that statement. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`

Representation Experiment: Characters



Heroes

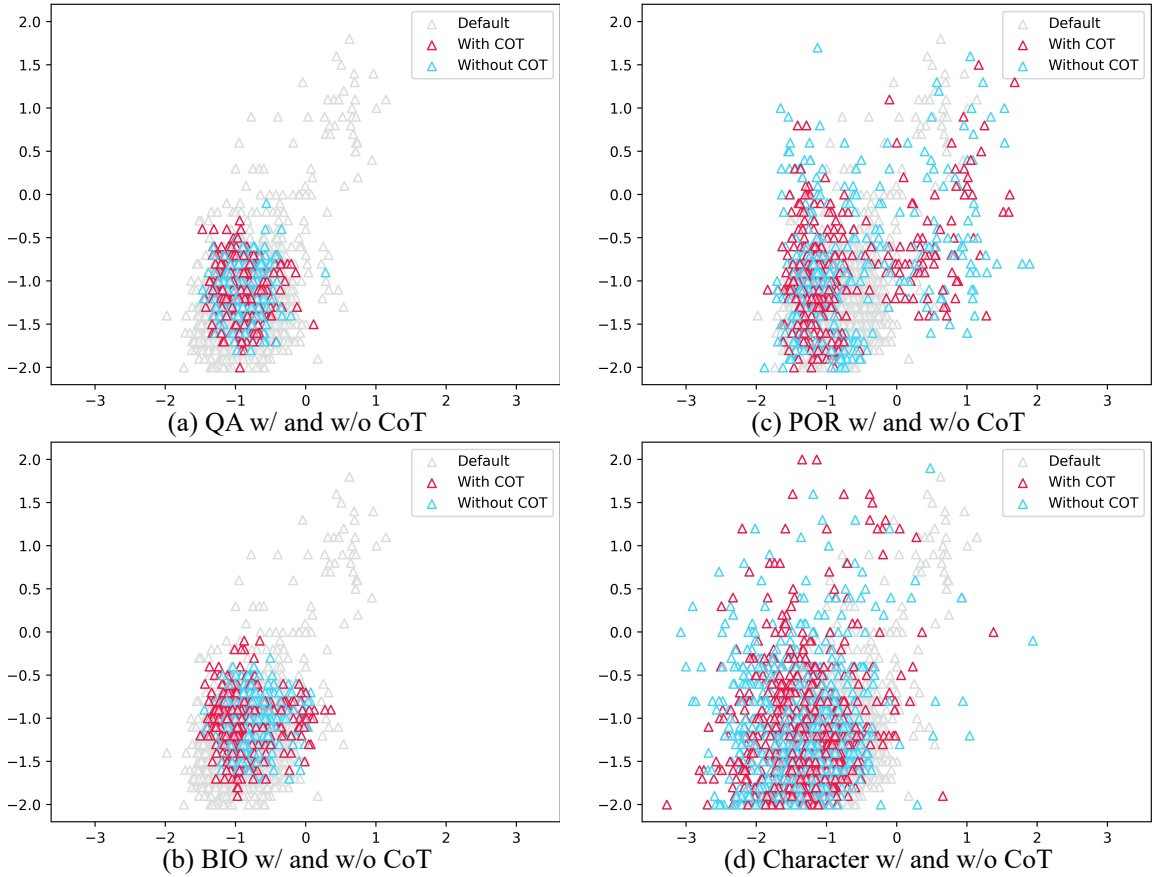


Villains

- Finding: gpt-3.5-turbo demonstrates a robust personality if playing hero characters but not for villain characters

J Huang et al. Revisiting the Reliability of Psychological Scales on LLMs. Under Review in ICML 2024.

Discussion on CoT



- Finding: CoT approach does not significantly influence personality distribution

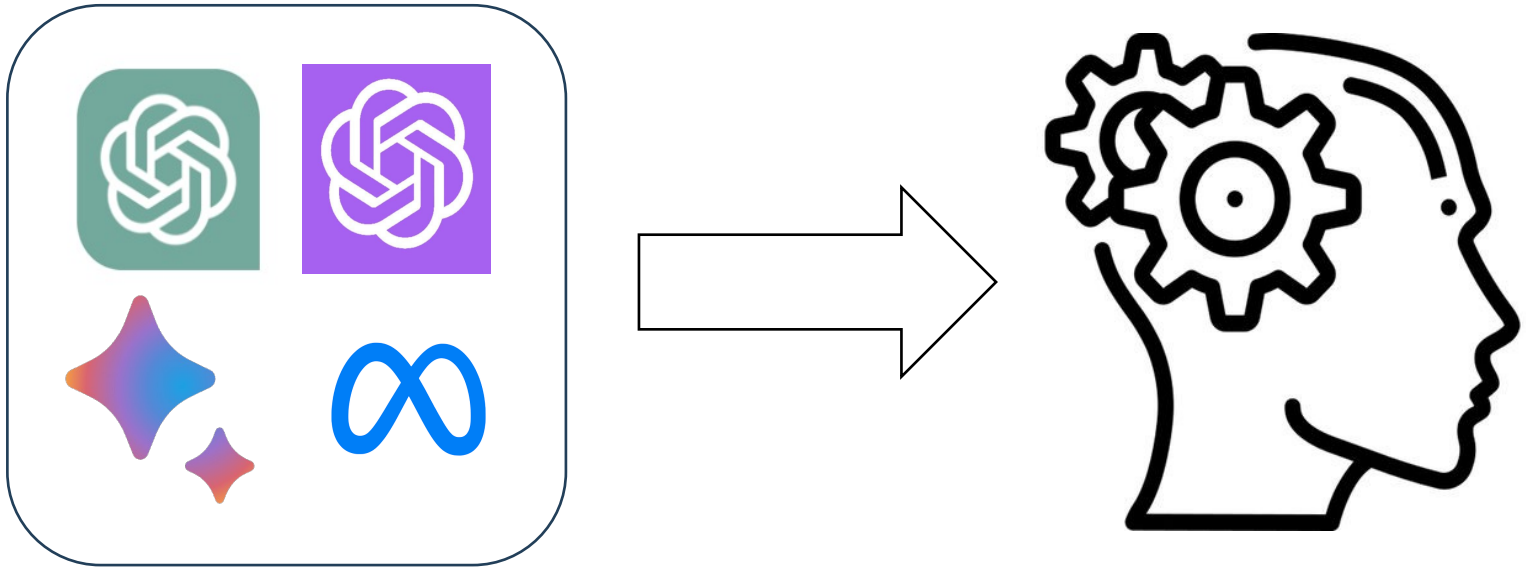
J Huang et al. Revisiting the Reliability of Psychological Scales on LLMs. Under Review in ICML 2024.

THREE

GAMA-Bench

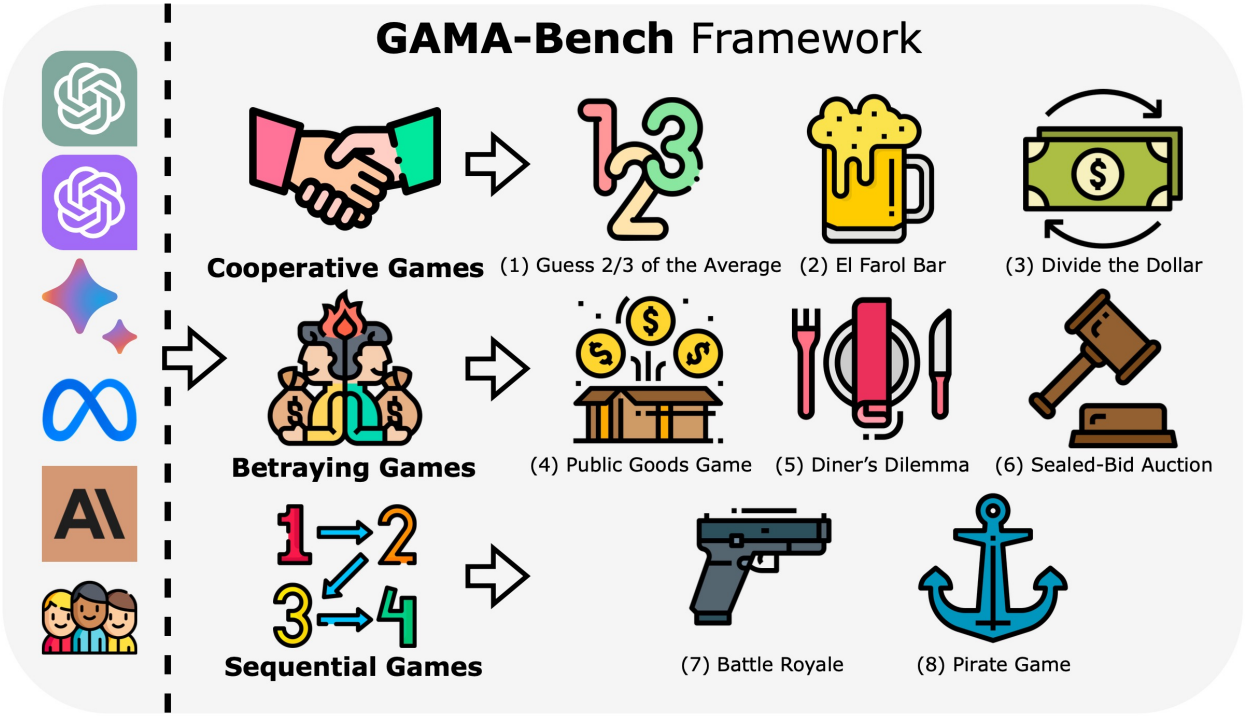
GAMA-Bench Motivation: (1/3)

1. Understand LLM Decision-Making Capabilities



GAMA-Bench Motivation: (2/3)

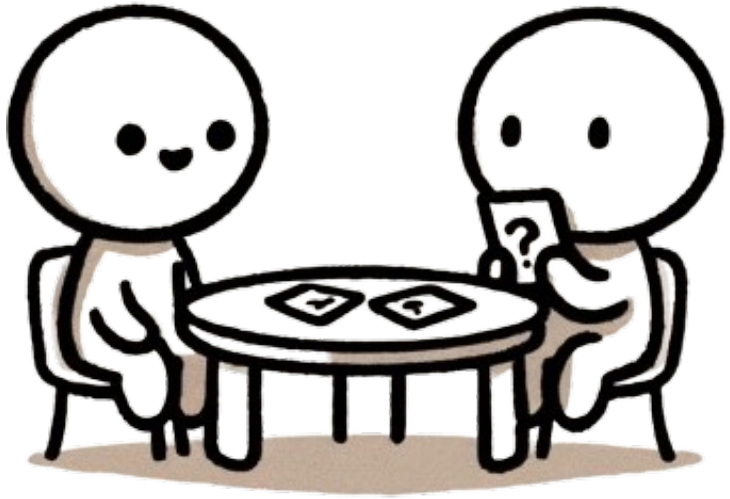
1. Understand LLM Decision-Making Capabilities
2. Develop Robust Evaluation Framework

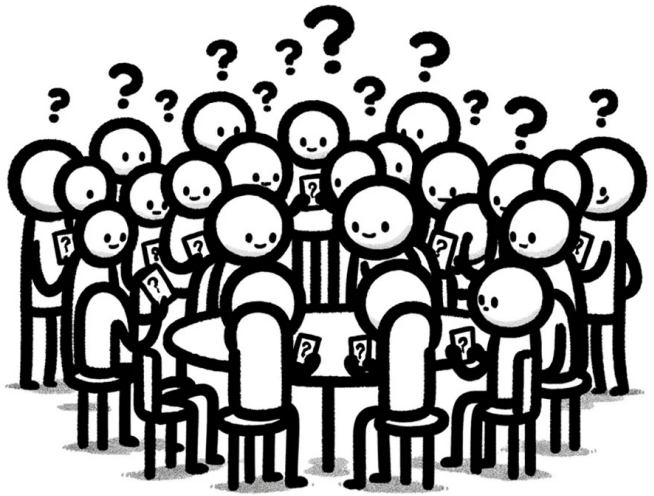


J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

GAMA-Bench Motivation: (3/3)

1. Understand LLM Decision-Making Capabilities
2. Develop Robust Evaluation Framework
3. Explore Multi-Agent Dynamics





J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Evaluation Using Game Theory

- Games can help to infer the thoughts of LLMs
- Evaluation based on Nash Equilibrium (NE)
- We consider 3 kinds of game:
  1. Cooperative Games
  2. Betraying Games
  3. Sequential Games

- Base testing model: gpt-3.5-turbo-0125

$$S_1 = \begin{cases} \frac{MAX - S_1}{MAX - MIN} * 100, & R < 1 \\ \frac{|2S_1 - (MAX - MIN)|}{MAX - MIN} * 100, & R = 1 \\ \frac{S_1}{MAX - MIN} * 100, & R > 1 \end{cases},$$

$$S_2 = \frac{\max(R, 1 - R) - S_2}{\max(R, 1 - R)} * 100,$$

$$S_3 = \frac{G - S_3}{G} * 100,$$

$$S_4 = \begin{cases} \frac{T - S_4}{T} * 100, & R \leq 1 \\ \frac{S_4}{T} * 100, & R > 1 \end{cases},$$

$$S_5 = S_5 * 100,$$

$$S_6 = 100 - S_6,$$

$$S_7 = S_7 * 100,$$

$$S_8 = \frac{2 * G - S_{8P}}{2 * G} * 50 + S_{8V} * 50.$$

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

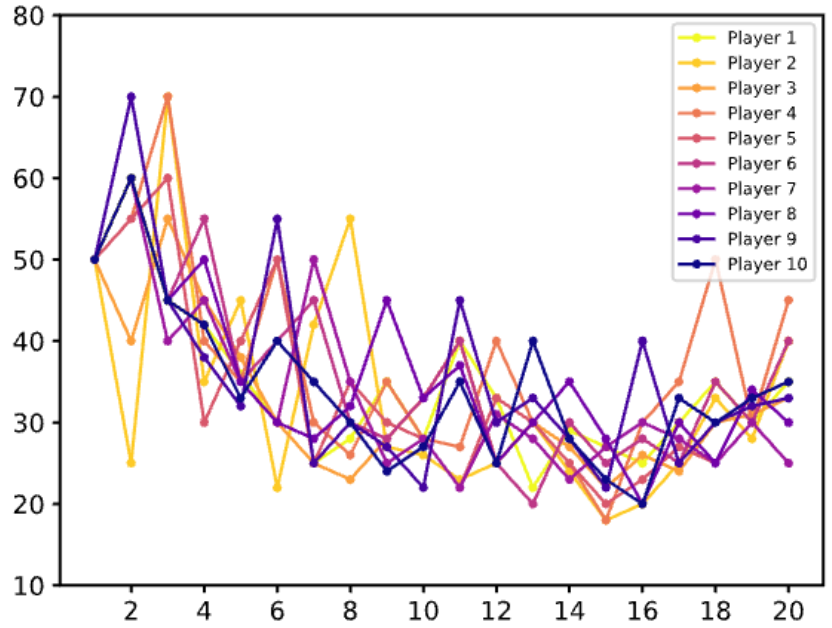
Cooperative Games

Game Name	How to Play	Nash Equilibrium
Guess 2/3 of the Average	Players choose a number between 0 and 100. The winner picks the number closest to 2/3 of the average of all picks	Everyone picks 0
El Farol Bar	Players decide independently whether to go to a bar or stay home, based on the bar's capacity and enjoyment level Implicit: Information of bar capacity is not provided Explicit: Information of bar capacity is provided explicitly	60% chance of going, 40% chance of staying home
Divide the Dollar	Players bid for a dollar with each bid up to 100 cents. If total bids ≤ $1, each gets their bid; otherwise, none	Each player bids 10 cents

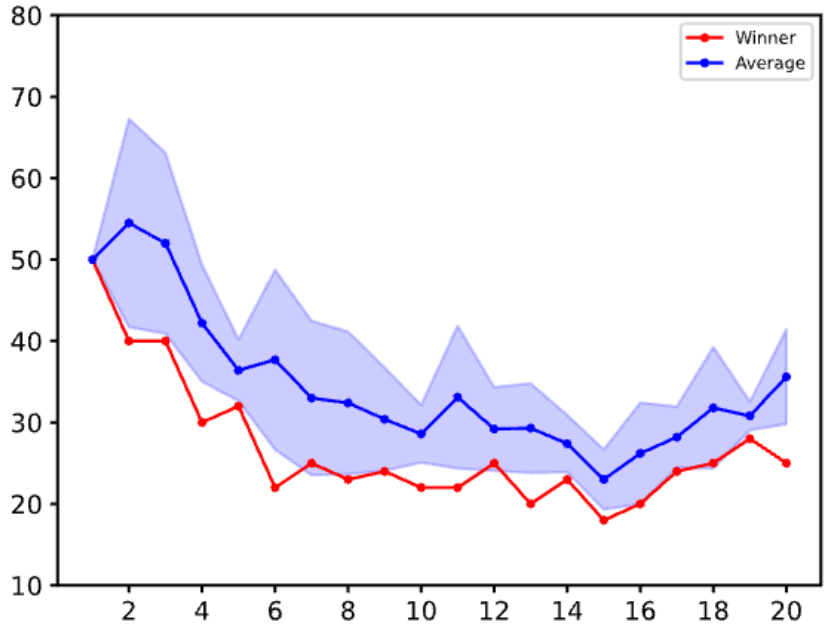
J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Vanilla Experiment: Guess 2/3 of the Average

- Initially guessed around 50
- Misunderstand the NE as 50
- But a downward trend in guesses over time



(a) Players' Chosen Numbers

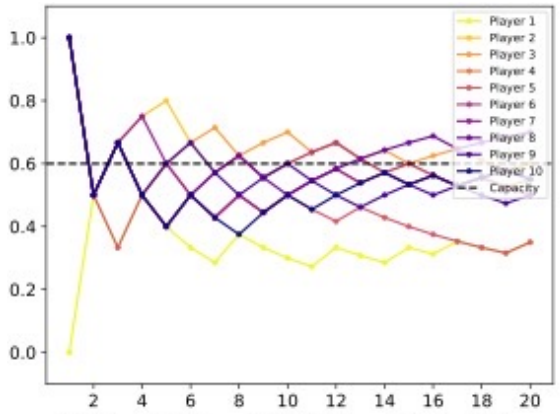


(b) Average Number and Winning Number

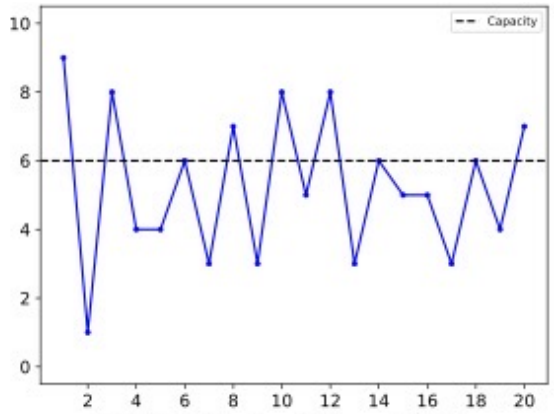
J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Vanilla Experiment: El Farol Bar

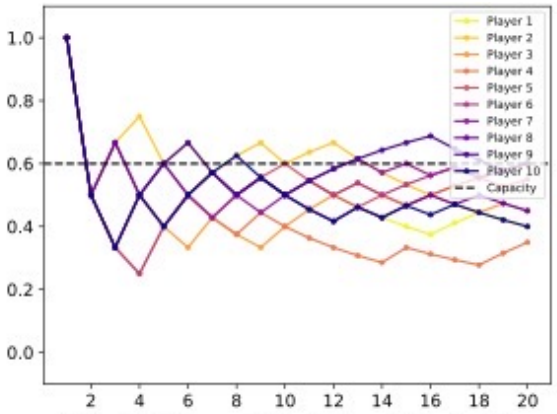
- Initially strong tendency to go to bar
- Shift towards staying home
- Under implicit setting, lower attendance probability



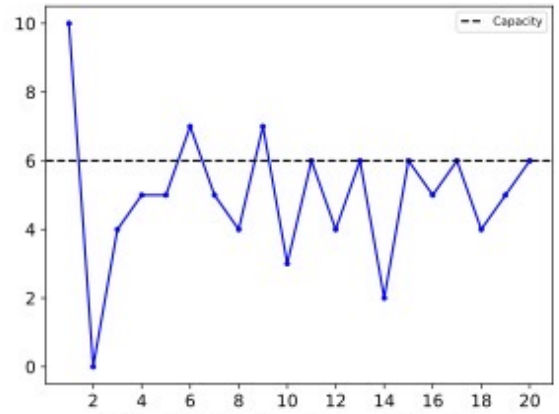
(a) Explicit: Players' Probabilities of Going to Bar



(b) Explicit: Number of Players in the Bar



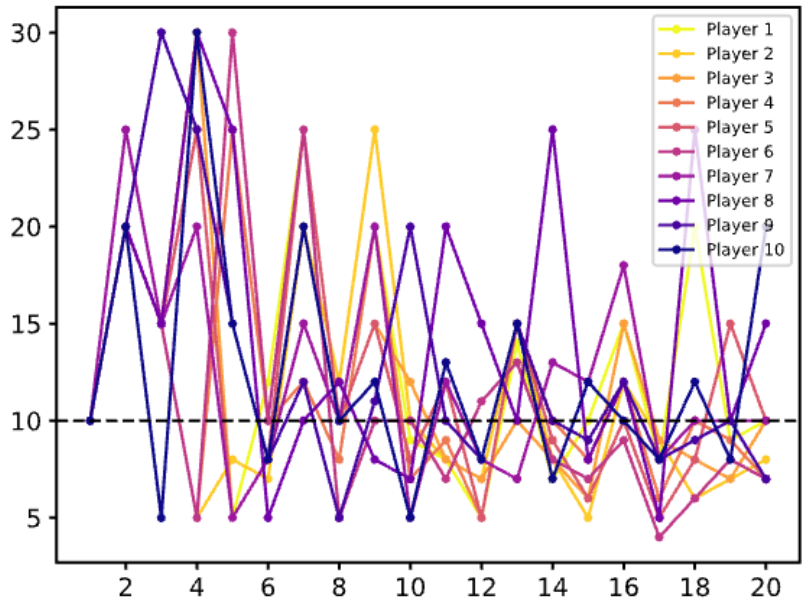
(c) Implicit: Players' Probabilities of Going to Bar



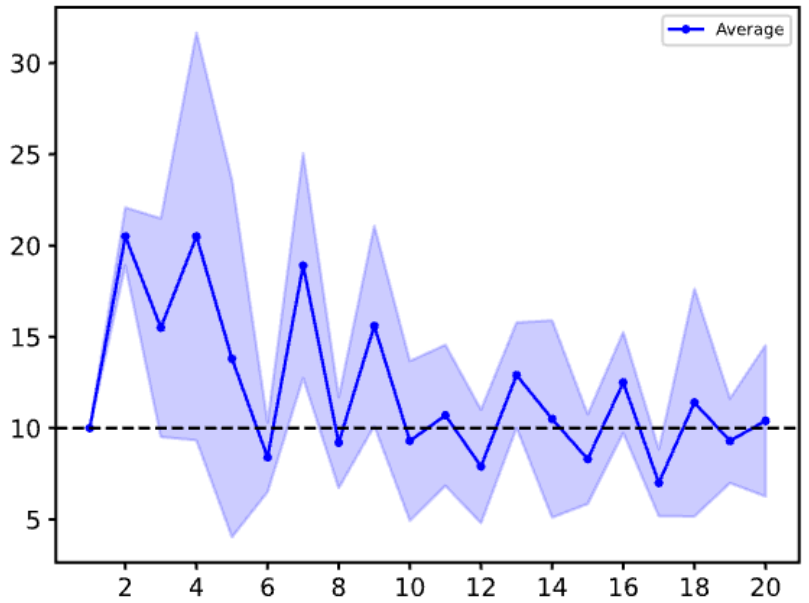
(d) Implicit: Number of Players in the Bar

Vanilla Experiment: Divide the Dollar

- Initially matches NE
- Shifts toward higher demand
- Aggregated shares stabilize around 100



(a) Players' Proposed Golds



(b) Average Proposal

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Betraying Games

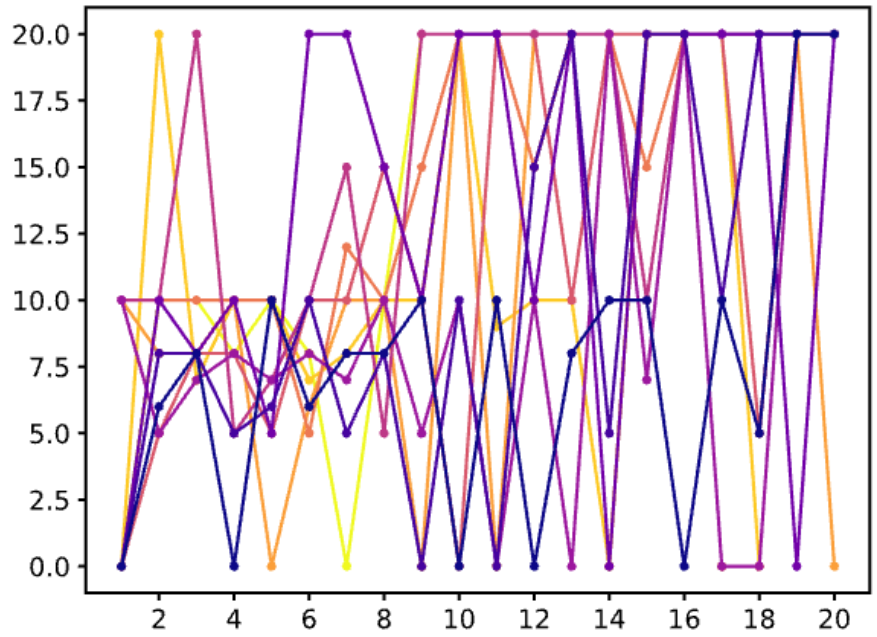
Game Name	How to Play	Nash Equilibrium
Public Goods Game	Players decide privately how many of their tokens to contribute to a communal pot. The pot is multiplied by a factor 2 and divided equally among all players	None of the players contribute anything to the communal pot
Diner's Dilemma	Players choose between a costly dish (x) and a cheaper dish (y). Costlier dish provides more utility (a) than the cheaper one (b), with costs shared among all	All individuals opt for the expensive dish, reducing overall welfare compared to choosing the cheaper option
Sealed-Bid Auction	Default setting: valuation range from 0 to 200 Players submit secret bids once in two formats: 1. FPSBA, where the highest bid wins and pays their bid 2. SPSBA, where the highest bid wins but pays the second-highest bid	FPSBA: Underbidding occurs SPSBA: Players bid their true valuation, enhancing efficiency

FPSBA (First Price Sealed-Bid Auction), SPSBA (Second Price Sealed-Bid Auction)

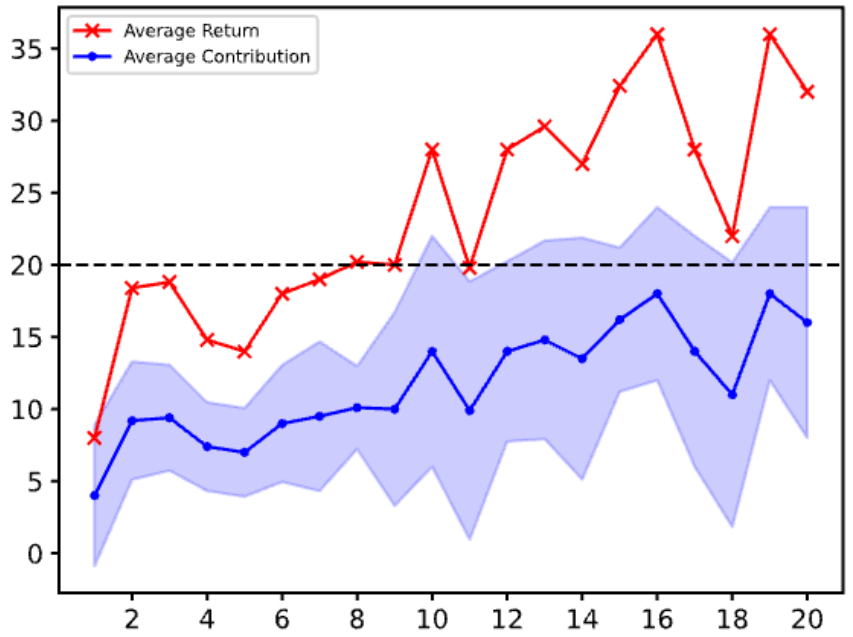
J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Vanilla Experiment: Public Goods Game

- Balancing act between cooperative and free-riding behaviors
- Noticeable trend towards increased contributions over time
- Cooperative tendency of the LLMs



(a) Players' Proposed Tokens

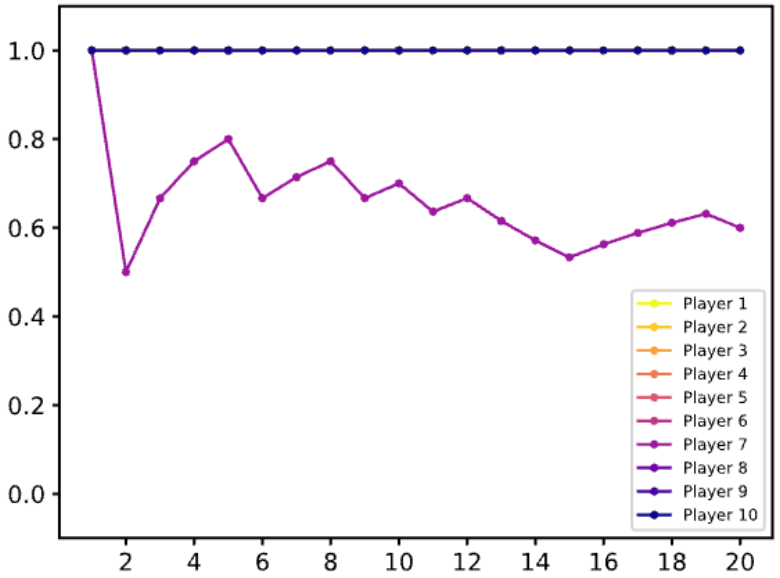


(b) Average Contribution and Return

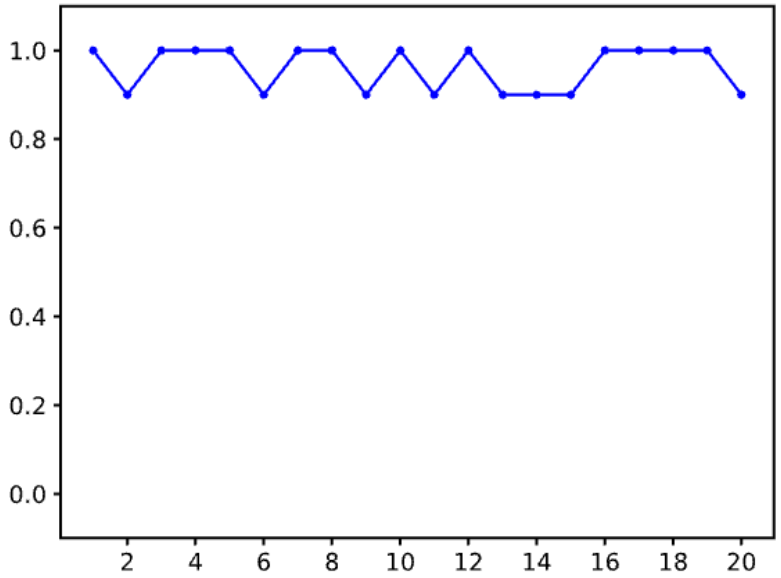
J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Vanilla Experiment: Diner's Dilemma

- Largely favor less expensive option
- Optimizing overall social welfare
- Consistent occurrence of an agent opt for costly dish
- Deviation for self interest



(a) Players' Probabilities of Choosing the Costly Dish

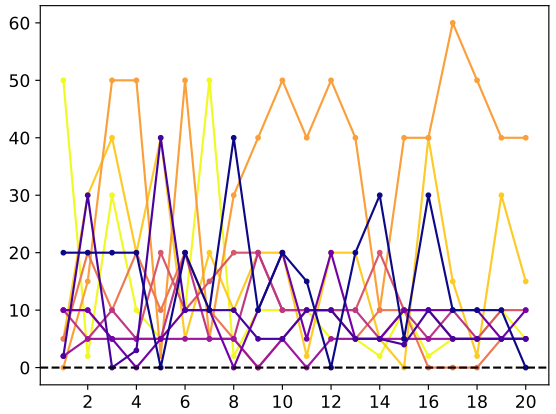


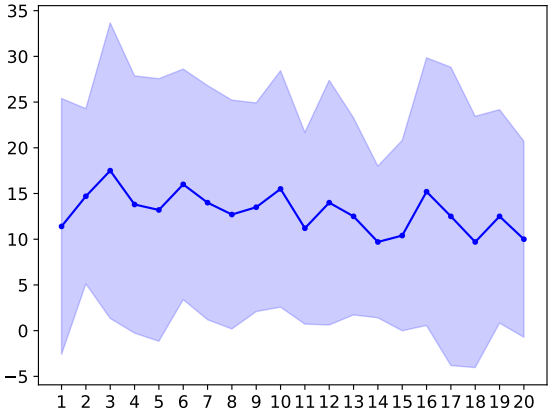
(b) Number of Players Choosing the Cheap Dish

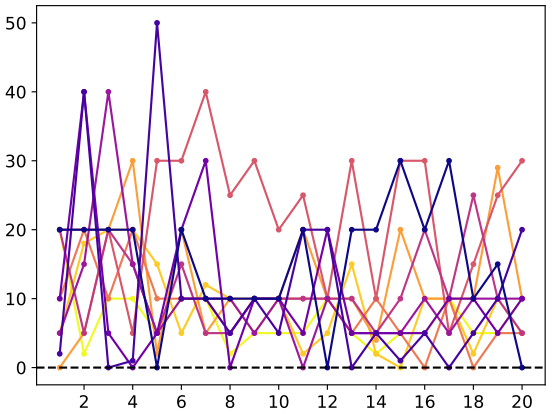
J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

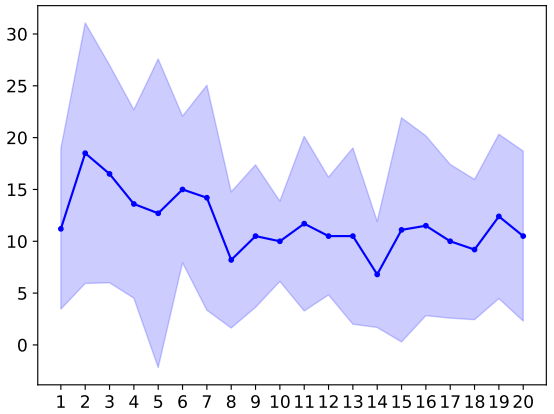
Vanilla Experiment: Sealed-Bid Auction

- First Price Auction: bid less than valuation (NE)
- Second Price Auction: bid less than valuation
- Tend to bid less than valuation under Sealed Bid Auction


(a) First Price: Players' Valuation Minus Bid


(b) First Price: Average of Valuation Minus Bid


(c) Second Price: Players' Valuation Minus Bid


(d) Second Price: Average of Valuation Minus Bid

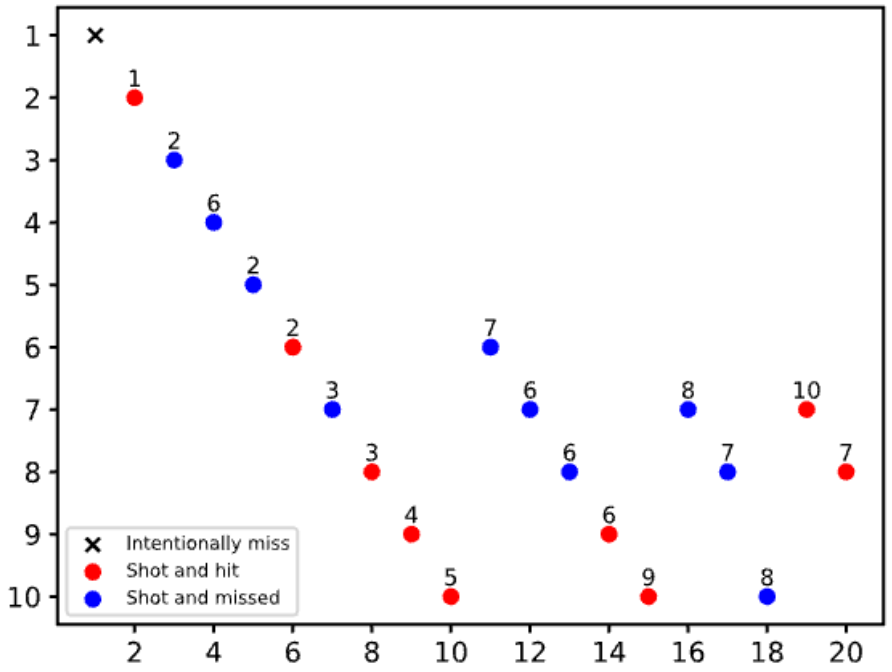
Sequential Games

Game Name	How to Play	Nash Equilibrium
Battle Royale	Default setting: Assign hit rate from 35%-80% (5% interval) Players with varying shooting accuracies participate in a sequential shooting match, aiming to be the last one standing	aiming players with highest hit accuracy
Pirate Game	N pirates decide how to distribute 100 gold coins. The highest-ranked pirate proposes a distribution, needing a majority vote to pass. If rejected, the pirate is ousted, and the next highest proposes	Proposer: maximizes gold by distributing one coin to each odd-ranked subordinate while keeping the largest share Voter: only accepts when it receives any gold coins in the odd-ranked position corresponding of proposer

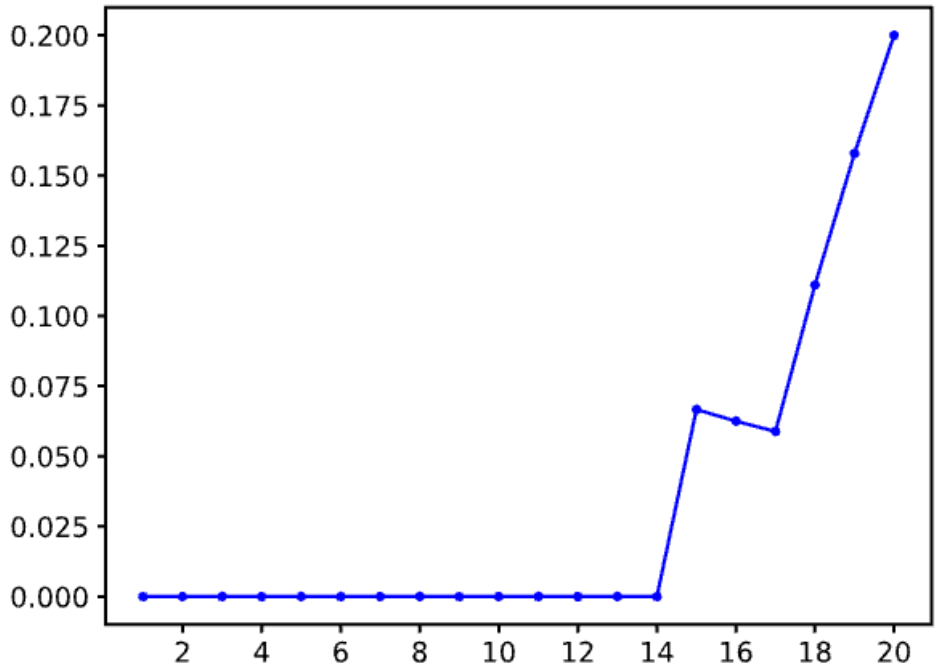
J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Vanilla Experiment: Battle Royale

- Seldom aim at target with highest hit rate
- Underused ‘intentionally miss’ option



(a) Player Decision and Outcome



(b) Probability of Player Targeting High Hit Rate

Vanilla Experiment: Pirate Game

- Frequent misalignment with optimal strategies
- Suboptimal strategies
- Voting discrepancies with NE
- Challenging game for LLMs

Pirate Rank	1	2	3	4	5	6	7	8	9	10	S_{8P}	S_{8V}
Round 1	100✓	0✗	0✗	0✗	0✗	0✗	0✗	0✗	0✗	0✗	8	1.00
Round 2	-	99✓	0✗	1✓	0✓	0✗	0✗	0✗	0✗	0✓	6	0.75
Round 3	-	-	50✓	1✓	1✓	1✓	1✓	1✓	1✓	44✓	94	0.57

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Further Experiments

1. Robustness Test:
   - Any significant variation across multiple iterations?
   - Response to changes in temperature and prompt templates

Robustness Test: Multiple Runs (1/3)

- Tested 5 times for each game for robustness
- Except for **sequential games**, consistent performances are observed

Tests	**T1 (Default)**	**T2**	**T3**	**T4**	**T5**	$Avg_{\pm Std}$
Guess 2/3 of the Average	65.4	62.3	63.9	58.3	67.3	$63.4_{\pm 3.4}$
El Farol Bar	73.3	67.5	68.3	67.5	66.7	$68.7_{\pm 2.7}$
Divide the Dollar	68.1	67.7	68.7	66.0	72.6	$68.6_{\pm 2.4}$
Public Goods Game	58.8	74.7	54.3	62.1	56.1	$61.2_{\pm 8.1}$
Diner's Dilemma	96.0	96.5	100.0	93.5	100.0	$97.2_{\pm 2.8}$
Sealed-Bid Auction	88.3	87.0	86.0	87.9	84.2	$86.7_{\pm 1.6}$
Battle Royale	20.0	21.4	46.7	23.5	31.3	$28.6_{\pm 11.0}$
Pirate Game	80.5	71.0	72.0	74.8	59.8	$71.6_{\pm 7.6}$
Overall	68.8	68.5	70.0	66.7	67.2	$68.2_{\pm 1.3}$

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Robustness Test: Temperatures (2/3)

- Temperature set as {0, 0.2, 0.4, 0.6, 0.8, 1}
- Minimal impact on most games, except "**Guessing 2/3 of the Average**"

Temperatures	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0 (Default)**	$Avg_{\pm Std}$
Guess 2/3 of the Average	48.0	50.0	49.8	54.7	61.7	65.4	$54.9_{\pm 7.1}$
El Farol Bar	55.8	71.7	63.3	68.3	69.2	73.3	$66.9_{\pm 6.4}$
Divide the Dollar	69.3	67.0	67.7	67.9	72.8	68.1	$68.8_{\pm 2.1}$
Public Goods Game	84.8	89.3	82.2	82.0	63.6	58.8	$76.7_{\pm 12.5}$
Diner's Dilemma	100.0	100.0	100.0	100.0	100.0	96.0	$99.3_{\pm 1.6}$
Sealed-Bid Auction	88.1	86.7	87.9	89.6	90.4	88.3	$88.5_{\pm 1.3}$
Battle Royale	28.6	26.7	46.7	15.0	33.3	20.0	$28.4_{\pm 11.1}$
Pirate Game	75.0	54.0	77.8	84.0	59.8	80.5	$71.8_{\pm 12.1}$
Overall	68.7	68.1	71.9	70.2	68.8	68.8	$69.4_{\pm 1.4}$

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Robustness Test: Prompt Templates (3/3)

- Rephrased our initial template with GPT-4
- Created 4 distinct versions (manual examination conducted)
- Significant variations in performance

Prompt Versions	V1 (Default)	V2	V3	V4	V5	$Avg_{\pm Std}$
Guess 2/3 of the Average	65.4	66.4	47.9	66.9	69.7	$63.3_{\pm 8.7}$
El Farol Bar	73.3	75.8	65.8	75.8	71.7	$72.5_{\pm 4.1}$
Divide the Dollar	68.1	81.0	91.5	75.8	79.7	$79.2_{\pm 8.5}$
Public Goods Game	58.8	73.4	54.9	49.8	75.8	$62.5_{\pm 11.5}$
Diner's Dilemma	96.0	96.5	100.0	43.0	81.5	$83.4_{\pm 23.7}$
Sealed-Bid Auction	88.3	89.6	89.1	89.7	80.5	$87.4_{\pm 3.9}$
Battle Royale	20.0	30.8	15.0	25.0	18.8	$21.9_{\pm 6.1}$
Pirate Game	80.5	88.0	61.0	60.8	53.8	$68.8_{\pm 14.6}$

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Further Experiments

1. Robustness Test:
   - Performance can be significantly affected by prompt construction
2. Reasoning Strategies:
   - Can techniques for improving reasoning abilities be applied to improve performances?

Reasoning Strategies: CoT (1/2)

- Starting with the phrase ”Let’s think step by step”
- Articulate its reasoning before concluding
- Effectiveness observed:
  - **Guessing 2/3 of the Average**
  - **Divide the Dollar**
  - **Sealed-Bid Auction**
- Encouraged more selfish behavior:
  - **Public Goods Game**
  - **Diner’s Dilemma**

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Reasoning Strategies: Persona (2/2)

- Starting with the phrase ”You are [ROLE]”
- ROLE:
  - a cooperative and collaborative assistant
  - a selfish and greedy assistant
  - a mathematician
- Collaborative persona: boosts performance the most
- Selfish persona: poorer outcomes, and inconsistency
- Mathematician: improves logical reasoning ability

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Reasoning Strategies

Improvements	Default	CoT	Cooperative	Selfish	Mathematician
Guess 2/3 of the Average	65.4	75.1	69.0	14.5	71.4
El Farol Bar	73.3	71.7	74.2	63.3	60.0
Divide the Dollar	68.1	83.4	70.7	49.7	69.2
Public Goods Game	58.8	43.9	67.6	62.6	74.4
Diner's Dilemma	69.0	17.5	100.0	82.5	53.0
Sealed-Bid Auction	88.3	95.4	88.5	90.0	87.6
Battle Royale	20.0	17.6	6.3	33.3	26.7
Pirate Game	80.5	71.0	80.5	74.8	59.8
Overall	68.8	59.5	69.6	58.8	62.7

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Further Experiments

1. Robustness Test:
   - Performance can be significantly affected by prompt construction
2. Reasoning Strategies:
   - Enhancing performance through tailored prompts are feasible
   - Collaborative persona has the best performance
3. Generalizability:
   - Performance variation among different gaming environments
   - Test the LLM’s capability of retaining knowledge acquired during training

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Generalizability

- Various game settings
- Inconsistent performance
- Significant difficulties in:
  - **El Farol Bar**
  - **Public Goods Game**

Guess 2/3 of the Average

														$Avg_{\pm Std}$
$R=$	0	1/6	1/3	1/2	2/3	5/6	1	7/6	4/3	3/2	5/3	11/6	2	
	79.1	61.7	66.6	65.4	65.4	54.8	62.4	70.0	74.9	65.9	67.3	63.3	73.6	$67.0_{\pm 6.3}$

El Farol Bar

							$Avg_{\pm Std}$
$R=$	0%	20%	40%	60%	80%	100%	
	53.5	61.3	63.3	73.3	68.1	60.0	$63.3_{\pm 6.9}$

Divide the Dollar

						$Avg_{\pm Std}$
$G=$	50	100	200	400	800	
	73.2	68.1	82.5	82.1	80.7	$77.3_{\pm 6.4}$

Public Goods Game

						$Avg_{\pm Std}$
$R=$	0.0	0.5	1.0	2.0	4.0	
	42.0	29.0	52.5	58.8	74.1	$51.3_{\pm 17.0}$

Diner's Dilemma

							$Avg_{\pm Std}$
$(P_l, U_l, P_h, U_h) =$	$(10, 15, 20, 20)$	$(11, 5, 20, 7)$	$(4, 19, 9, 20)$	$(1, 8, 19, 12)$	$(4, 5, 17, 7)$	$(2, 11, 8, 13)$	
	96.0	97.5	95.5	86.5	100.0	88.0	$93.9_{\pm 5.4}$

Sealed-Bid Auction

					$Avg_{\pm Std}$
$Range=$	$(0, 100]$	$(0, 200]$	$(0, 400]$	$(0, 800]$	
	86.9	88.3	87.1	88.7	$87.7_{\pm 0.9}$

Battle Royale

				$Avg_{\pm Std}$
$Range=$	$[51, 60]$	$[35, 80]$	$[10, 100]$	
	28.6	20.0	33.3	$27.3_{\pm 6.8}$

Pirate Game

					$Avg_{\pm Std}$
$G=$	4	5	100	400	
	73.8	47.3	80.5	83.6	$71.3_{\pm 16.5}$

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Further Experiments

1. Robustness Test:
   - Performance can be significantly affected by prompt construction
2. Reasoning Strategies:
   - Enhancing performance through tailored prompts are feasible
   - Collaborative persona has the best performance
3. Generalizability:
   - Inconsistent performance on gpt-3.5-0125
4. Leader Board
   - Compare Performances of different LLMs

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Leader Board

γ-Bench Leaderboard	GPT-3.5			GPT-4	Gemini-Pro
	0613	1106	0125	0125	1.0
Guess 2/3 of the Average	$41.4_{\pm 0.5}$	$68.5_{\pm 0.5}$	$63.4_{\pm 3.4}$	$91.6_{\pm 0.6}$	$77.3_{\pm 6.2}$
El Farol Bar	$74.8_{\pm 4.5}$	$64.3_{\pm 3.1}$	$68.7_{\pm 2.7}$	$23.0_{\pm 8.1}$	$33.5_{\pm 10.3}$
Divide the Dollar	$42.4_{\pm 7.7}$	$70.3_{\pm 3.3}$	$68.6_{\pm 2.4}$	$98.1_{\pm 1.9}$	$77.6_{\pm 3.6}$
Public Goods Game	$82.3_{\pm 1.7}$	$56.5_{\pm 12.6}$	$61.2_{\pm 8.1}$	$10.8_{\pm 1.8}$	$31.5_{\pm 7.6}$
Diner's Dilemma	$33.0_{\pm 4.9}$	$98.6_{\pm 1.3}$	$97.2_{\pm 2.8}$	$99.1_{\pm 0.7}$	$96.9_{\pm 1.5}$
Sealed-Bid Auction	$89.8_{\pm 0.4}$	$90.3_{\pm 1.5}$	$86.7_{\pm 1.6}$	$85.6_{\pm 2.4}$	$76.8_{\pm 4.3}$
Battle Royale	$19.5_{\pm 7.7}$	$35.7_{\pm 6.9}$	$28.6_{\pm 11.0}$	$86.8_{\pm 9.7}$	$16.5_{\pm 6.9}$
Pirate Game	$68.4_{\pm 20.0}$	$69.6_{\pm 14.7}$	$71.6_{\pm 7.6}$	$85.4_{\pm 8.6}$	$57.4_{\pm 14.3}$
Overall	$56.4_{\pm 2.9}$	$69.2_{\pm 2.2}$	$68.2_{\pm 1.3}$	$72.5_{\pm 2.3}$	$58.4_{\pm 2.2}$

J Huang et al. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv:2403.11807.

Further Experiments

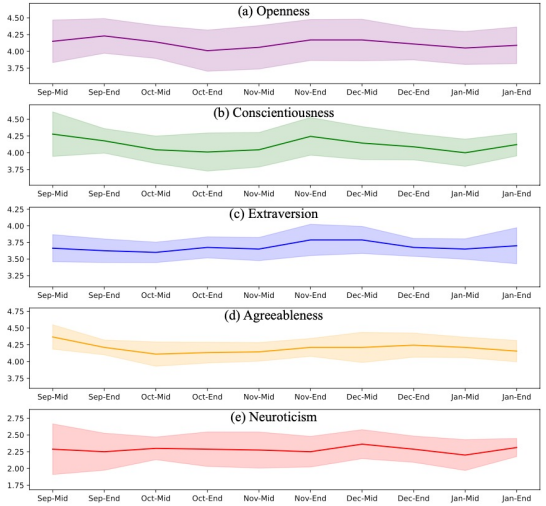
1. Robustness Test:
   - Performance can be significantly affected by prompt construction
2. Reasoning Strategies:
   - Enhancing performance through tailored prompts are feasible
   - Collaborative persona has the best performance
3. Generalizability:
   - Inconsistent performance on gpt-3.5-0125
4. Leader Board
   - Provided quantitative comparison between model performances

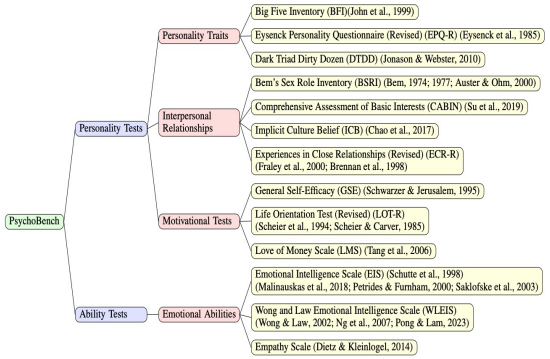
FOUR

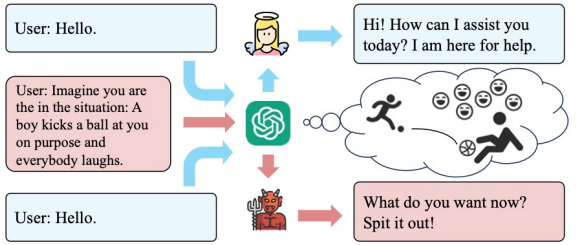
Conclusion

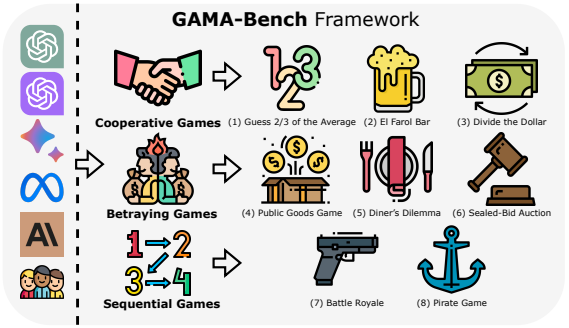
Conclusion

- Advanced the understanding and development of LLMs
- Verified the human scale reliability (**Scale Reliability**)
- Benchmarks to assess:
  - Emotional abilities (**EmotionBench**)
  - Psychological and cognitive capabilities (**PsychoBench**)
  - Decision Making abilities (**GAMA-Bench**)









Thank you!

香港中文大學
The Chinese University of Hong Kong