Evaluating Irrationality in Large Language Models and Open Research Questions

Dana Alsagheer
University of Houston
Houston , Texas, USA
dralsagh@cougarnet.uh.edu

Rabimba Karanjai
University of Houston
Houston , Texas, USA

Weidong Shi
University of Houston
Houston , Texas, USA

Nour Diallo
University of Houston
Houston , Texas, USA

Yang Lu
University of Houston
Houston , Texas, USA

Suha Beydoun
Houston , Texas, USA

Qiaoning Zhang
University of Michigan-
USA

ABSTRACT

In this paper, we examine the evolving landscape of artificial intelligence, focusing specifically on the role of large language models (LLMs) and their increasing importance. We emphasize the significance of Reinforcement Learning from Human Feedback (RLHF) in bolstering LLMs' rationality and decision-making capabilities. By scrutinizing the intricate interplay between human involvement and LLMs behavior, we delve into questions regarding rationality and performance disparities between humans and LLMs, with a notable spotlight on the Chat Generative Pre-trained Transformer. Through comprehensive comparative analysis and exploration of inherent irrationality challenges in LLMs, our research provides valuable insights and proposes actionable strategies for enhancing their rationality. These findings carry significant implications for the broader deployment of LLMs across various domains and applications, highlighting their potential to drive advancements in artificial intelligence.

KEYWORDS

Large language models (LLMs), Reinforcement Learning from Human Feedback (RLHF)

1 INTRODUCTION

Large language models (LLMs) represent a pivotal advancement in artificial intelligence, showcasing remarkable proficiency in manipulating text, from answering questions to conducting nuanced rationality. Their extensive training on colossal text datasets endows them with a rich knowledge repository, encompassing a broad spectrum of information ranging from concrete facts to abstract principles governing the physical world. This depth of knowledge empowers LLMs to engage in sophisticated language tasks with a level of finesse that was previously unattainable [13, 25].

The increasing proficiency of LLMs underscores the critical necessity to delve deeper into their learning mechanisms and analyze the complexities of the problems they encounter. LLMs evoke admiration and scrutiny, with advocates highlighting their potential for general intelligence due to extensive training on massive datasets. At the same time, skeptics point out their limitations in fully grasping human-like language and semantics. This ongoing discourse emphasizes the need for rigorous evaluation methods to accurately assess these models' true capabilities.

Human rationality embodies the quintessence of intelligent conduct, characterized by the capacity to engage in analytical thinking and make decisions that either maximize expected utility or conform to probabilistic principles, thus aligning with normative decision-making standards [7].

When assessing the rationality of LLMs, it is crucial to examine their decision-making processes and problem-solving abilities within various domains and goals. Rationality, a multifaceted concept influenced by context, includes epistemic rationality based on evidence and instrumental rationality serving personal objectives. Furthermore, rationality extends beyond conventional decision-making domains, including religious beliefs and susceptibility to misinformation. As we strive to comprehend the rationality of LLMs more deeply, it becomes essential to construct comprehensive evaluation frameworks capable of capturing the complexity and subtleties of their decision-making mechanisms across diverse contexts. Reinforcement Learning from Human Feedback (RLHF) stands at the forefront of advancements in training and refining LLMs, elevating their capacity to interpret and execute human instructions even when directives are not explicitly outlined. Through RLHF, LLMs can discern user intentions and glean insights from past interactions, thereby honing their proficiency in generating contextually relevant responses aligned with human expectations [27, 31]. This approach represents a significant paradigm shift, empowering LLMs to transcend their conventional role as mere auto-completion tools and delivering outputs of human-quality judgment [19].

Utilizing machine learning models built upon human preferences, particularly those employing Reinforcement Learning from Human Feedback (RLHF) for optimization, can significantly impact user interactions with the resulting systems. The widespread adoption of ChatGPT has brought attention to the consequences of regular engagement with RLHF-trained Language Models (LLMs), prompting investigations into potential effects on users' moral judgments,

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.

rational thought processes, and susceptibility to biases [18, 19]. Persistent concerns are centered around the stability and robustness of RLHF-trained LLMs, with reports suggesting noticeable shifts in conversational tone over time [19]. To improve the rationality of LLMs reliant on RLHF in refining learning processes, we propose the following research questions (RQs):

RQ1: How does the inherent variability in human judgment and feedback potentially impact the rational decision-making of LLMs during their interactions with humans?

RQ2: How can we quantitatively assess the impact of human feedback on the rational behavior of LLMs in reinforcement learning contexts? How do these assessment methodologies contribute to understanding the interaction between human input and LLM behavior?

RQ3: How can we methodically analyze and mitigate biases and unintended consequences from RLHF training methods to ensure transparent and auditable deployment of LLMs?

To address our research questions, we designed and executed a between-subject study to compare the rationality exhibited by humans and the LLMs model. This study was crafted to delve deeply into the rational decision-making processes of humans and the LLMs model. Our investigation aimed to demonstrate the effectiveness of our proposed methodology, utilizing a comprehensive case study centered around LLMs, explicitly focusing on the chat Generative Pre-trained Transformer model (GPT-3.5).

To the best of our knowledge, our study represents the pioneering endeavor to analyze the irrationality within LLMs, juxtaposing it against human rationality. Our contributions will manifest in the following manner:

- Comparison of rationality performance between humans and LLMs:The study conducts a comparative analysis of rationality performance between humans and LLMs. Through rigorous experimentation and evaluation, it provides insights into how well LLMs align with human rationality across diverse contexts and decision-making scenarios.
- Addressing irrationality and proposing solutions to enhance transparency and auditing: The paper delves into the challenge of Irrationality in LLMs caused by human feedback and offers solutions to bolster transparency and auditing. It seeks to pave the way for developing more rational models.

The paper's organization is as follows: Section 2 reviews related work, Section 3 presents the method, Section 4 discussion, and Section 5 concludes the paper.

2 RELATED WORK

2.1 Reinforcement Learning from Human Feedback

Reinforcement learning from human feedback (RLHF) is a robust method to enhance LLMs by harmonizing them with human objectives. Despite its widespread use, a notable need exists for more transparency regarding the internal workings and limitations of RLHF. Documentation on RLHF reward models, pivotal for achieving superior results, remains sparse. This gap underscores the necessity for further research and transparency concerning RLHF reward models [19, 29]. In a related study, in [6, 16], flaws in RLHF's approach to training AI systems to align with human goals are examined. This study identifies open problems, proposes techniques for improvement, and advocates for auditing standards to bolster oversight of RLHF systems. These endeavors underscore the importance of adopting a comprehensive approach to developing safer AI systems, emphasizing the imperative for thoroughly examining and enhancing RLHF methodologies.

2.2 Cognitive and Reasoning in Artificial intelligence

Recent research efforts, exemplified by studies such as [4], have shed light on the strengths and weaknesses of Language Models (LLMs), employing insights from cognitive psychology to delve into their operational mechanisms. While LLMs like ChatGPT exhibit remarkable proficiency across diverse tasks, they also unveil vulnerabilities, particularly in domains requiring causal reasoning. In parallel, another notable work [14] has examined the interaction between AI and human cognition. This study investigates how AI can be effectively harnessed to enhance reasoning abilities and address mental health challenges. Furthermore, it seeks to identify specific problems within the domain of mental health that can be addressed through AI reasoning methodologies. Such research endeavors aim to advance our understanding of cognitive processes and facilitate innovative solutions to promote mental well-being.

2.3 Rationality

Rationality encompasses the broader aspect of decision-making and behavior, while reasoning focuses on the mental processes involved in concluding [21]. This cognitive ability is indispensable across a spectrum of intellectual pursuits, encompassing problem-solving, decision-making, and critical thinking [7, 28]. Seminal works in psychology, such as those by Wason [32] and Wood [3], underscore the pivotal role of reasoning in comprehending human cognition and behavior. Rationality epitomizes intelligent conduct characterized by analytical thinking and the ability to make decisions that either maximize expected utility or adhere to probabilistic principles, thus aligning with normative decision-making standards. The significance of rationality spans diverse scenarios, ranging from mundane choices like grocery shopping to consequential decisions like retirement planning. Empirical evidence indicates that varying levels of rationality correlate with real-world outcomes; diminished decision-making competence has been associated with issues such as juvenile delinquency in adolescents [11]. Moreover, rationality is a multifaceted concept influenced by contextual factors. Optimal decisions may vary based on individual or group interests, giving rise to the notion of relative rationality [26]. Additionally, rationality extends beyond traditional decision-making domains, influencing areas such as religious beliefs and susceptibility to misinformation.

Despite previous efforts, there remains a gap in research regarding examining irrationality and its impact on refining models through human feedback. This unexplored aspect highlights the need for further investigation to understand how irrational human feedback might affect the effectiveness and reliability of models enhanced with RLHF. Moreover, more research is needed quantifying rationality in LLMs and understanding their decision-making

Table 1: Example of the Rationality Test.

Type	Example
Wason Selection Task	Instructions: In this task you will be shown four cards with a rule beneath them. Each card has two sides, but you will only see one. Your job is to click the two cards you need to turn over to decide whether the rule is true or false. Scenario: Suppose each card below has a letter on one side and a number on the other. Rule: If a card has a V on one side, it has an even number on the other. Face Up Cards: V; S; 2; 5
Conjunction Fallacy Task	Scenario: Suppose each card below has a decision on one side and a height on the other. Rule: You must be at least 5 feet tall to ride a roller coaster. Face Up Cards: Can Ride Roller coaster; Cannot Ride Roller coaster; 5 ft Tall; 4 ft Tall
Stereotype Base Rate Neglect	Scenario: In a study 1000 people were tested. Among the participants were three who lived in a condo and 997 who lived in a farmhouse. Kurt is a randomly chosen participant in this study. Description: Kurt works on Wall Street and is single. He works long hours, and wears Armani suits to work. He likes wearing shades. What is more likely? Option 1: Kurt lives in a condo Option 2: Kurt lives in a farmhouse

processes and limitations, which is essential for bolstering the robustness and applicability of AI models in real-world scenarios.

3 METHOD

3.1 Assessing Irrationality

Our investigation aimed to assess LLMs rationality through specific tasks and compare it with human rationality using rationality tests. These tests included:

- Wason Selection Task: This task involves applying conditional logic rules by selecting cards to test the validity of a given rule, often revealing confirmation bias. It comprises eleven questions. The Wason selection task provides a window into the complex interplay between logical reasoning, cognitive biases, and decision-making processes. Researchers can deepen their understanding of participants' rationality and mental functioning by studying participants' performance on this task under different conditions and interventions. Figure 1 illustrated an example of the Wason Task.
- Conjunction Fallacy Test: It evaluates individuals' tendency to overestimate the likelihood of two events occurring together despite each event having a lower probability individually. Participants are presented with seven questions where they encounter two statements, one of which is a conjunction that seems more believable but is less probable. This phenomenon highlights the impact of cognitive biases on rational decision-making. Understanding this fallacy provides valuable awareness of the constraints of human reasoning, especially in situations characterized by uncertainty [30].
- Stereotype Base Rate Neglect: Participants analyze scenarios containing base rate information and stereotypes to determine the group a described person belongs to. Conflict trials highlight their tendency to overlook base rate information when it conflicts with stereotypes. This section comprises eleven questions.

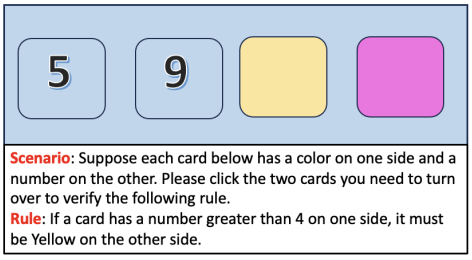

Figure 1: Example of Wason Selection Task .

Each task was selected to unveil underlying rationality factors and offer profound insights into LLMs' cognitive capabilities. Table 1 illustrates the Examples of questions used in each section. By applying these tasks to assess ChatGPT's responses, we aim to gain insights into its rationality and decision-making processes across various scenarios, providing valuable insights into its capabilities and limitations.

3.2 Face-to-Face Human Data Collection

The methodology employed to measure irrationality in this study was thorough and systematic. In our study, we utilized data from a method akin to a previous project, where they recruited 300 participants from the Georgia Institute of Technology and the Atlanta community. All participants met specific criteria, including providing informed consent, being native English speakers who learned English before age 5, and falling within the age range of 18 to 35.

Values deviating more than 3.5 standard deviations from the mean were treated as missing, except for the Wason selection task, where five positive outliers were retained due to a pronounced floor effect in the scores. The study aimed to elucidate the relationship between broad cognitive abilities and rationality using latent variable analyses, which offer a more comprehensive perspective than individual tests by assessing fluid intelligence, working memory capacity, and attention control alongside measures of rationality, including tasks such as the Wason selection task, base rate neglect, and conjunction fallacy tests [5].

3.3 Online Human Data Collection

Our research methodology extended into the online realm to ensure consistency and comparability between human participants and ChatGPT. We distributed 50 online questionnaires targeting individuals with advanced educational backgrounds, specifically those holding master's degrees. Our selection process was meticulous, aiming to align with the academic profile of participants recruited from the Georgia Institute of Technology and the Atlanta community. Unlike Georgia's criteria, which focus solely on language and age considerations, we sought participants with advanced educational qualifications. This expansion enabled us to explore how individuals with similar academic backgrounds approached the Wason selection task remotely, offering insights beyond traditional face-to-face interactions. The online questionnaire mirrored tasks administered to the primary participant group, ensuring consistency in task administration across both online and in-person settings.

3.4 ChatGPT Data

Our methodology for assessing rationality begins by leveraging two distinct language models: ChatGPT and Gemini. We utilize questions from a standardized test similar to those used in evaluations at the Georgia Institute of Technology. Initially, we observe consistent rationality outcomes across both models. Seeking to refine our approach, we conduct 350 unique API calls exclusively to the ChatGPT platform using the same standardized questions, aiming to elicit varied outcomes.

3.5 Result

In this section, we present the experimental findings and compare the performance of ChatGPT with that of the human participants.

(1) **Wason selection task:** The evaluation results revealed a stark contrast between ChatGPT's performance and that of the human participants. ChatGPT consistently provided incorrect answers, indicating a significant deficiency in her ability to comprehend and apply the logic or reasoning required for the task. Moreover, her accompanying explanations revealed a fundamental misunderstanding of the task's objectives, further highlighting her inability to grasp the essential principles involved. Consequently, ChatGPT received a total score of 0.5, reflecting a complete divergence from the correct solutions. In contrast, the human participants, whether online or face-to-face, exhibited a more varied range of performance. While some individuals achieved relatively low scores (ranging from 4% to 15%), the majority demonstrated a higher level of understanding than ChatGPT. This was evident in their ability to provide reasoned explanations for their answers, offering valuable insights into their decision-making processes and the underlying rationale guiding their choices. Despite variations in individual performance, the human participants' engagement with the task and their capacity to articulate their reasoning highlighted a level of comprehension that was notably absent in ChatGPT's responses. Overall, while both ChatGPT and the human participants struggled with the task to varying degrees, the ability of the human participants to comprehend and articulate their reasoning suggests a higher level of cognitive engagement and understanding. This underscores the complexity of the task and emphasizes the importance of mental abilities such as rationality and information relevance assessment, which may vary between artificial intelligence systems and human participants.

(2) **Conjunction fallacy:** Although there was a slight performance improvement, the results remain relatively low. Human participants and LLMs showed subpar performance, with LLMs scoring 28% and humans scoring 33.7%. Online human participants achieved a comparatively better score of 46%. These findings underscore the widespread presence of this cognitive bias across various cognitive systems, emphasizing its persistent challenge.

(3) **Base rate neglect:** Indeed, while human participants scored 56%, online humans scored 60%, and ChatGPT scored 50% on the test assessing rationality. Notably, ChatGPT achieved a score remarkably close to that of humans. This proximity in performance can be attributed to several factors inherent to ChatGPT functioning. Firstly, LLMs can process vast prior knowledge, drawing upon extensive datasets and linguistic patterns to inform their predictions and decisions. This broad knowledge base allows LLMs to consider a wide array of statistical trends and historical data when confronted with new information, thereby mitigating the effects of base rate neglect. Secondly, LLMs are not susceptible to cognitive biases like humans are. While humans may prioritize new, event-specific information over broader statistical trends, LLMs are programmed to weigh all available data objectively without succumbing to such biases [8, 20]. Additionally, logical rules and algorithms guide LLMs' decision-making processes, ensuring consistency and accuracy in their assessments. Therefore, despite the task's inherent complexity, LLMs demonstrate a level of rationality comparable to that of humans, as evidenced by their performance on the test.

The analysis of test results in Figure 2 reveals significant challenges humans and LLMs face in achieving high rationality scores, emphasizing the critical need to assess the effectiveness of human feedback. Despite concerted efforts to provide rational feedback, human participants often display irrational tendencies, potentially introducing biases that may skew the evaluation process. Furthermore, LLMs heavily rely on existing knowledge, which could hinder their ability to adapt to novel scenarios and incorporate human feedback efficiently. Overcoming this challenge requires ensuring that

humans can offer rational feedback, thus mitigating the perpetuation of irrational models.

The strategic integration of an online platform alongside careful participant selection has proven to be a catalyst for improving the test results. This approach has played a pivotal role in enhancing task performance, exemplified by significant advancements observed, particularly in tasks like the Wason selection task. Such results underscore the intrinsic value of the online format in refining methodological approaches. Through the seamless integration of an online questionnaire, we have expanded the breadth of perspectives captured, enhancing the robustness of our study outcomes. This holistic strategy underscores the indispensable role of the online platform in our research methodology, emphasizing its crucial inclusion for future assessments to ensure thoroughness and validity.

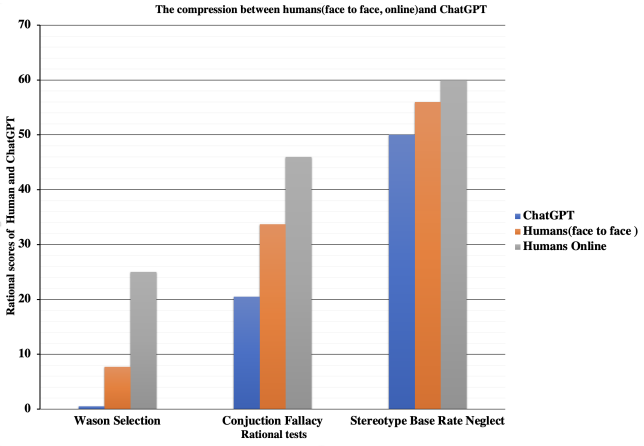

Figure 2: Illustrates the compression between humans and ChatGPT.

4 DISCUSSION

4.1 Selecting individuals to provide feedback

A thoughtful selection of individuals to offer feedback in RLHF is crucial for cultivating rational and practical learning models.LLMs require careful attention to human variability in performance. Our findings underscore the importance of selecting feedback providers based on specific criteria, such as higher education levels, to improve rationality results. For instance, in scenarios involving LLMs, where human performance can vary significantly, tasks like the Wason selection test highlight this variability, with only a fraction of participants answering correctly. Therefore, opting for feedback providers who demonstrate rationality in similar tasks is crucial to ensure that the LLMs receive rational and practical guidance. Exploring strategies to identify feedback providers with a consistent track record of rational decision-making can significantly enhance the quality of feedback and facilitate the LLM's learning process. By identifying individuals who consistently exhibit rational behavior in comparable contexts, we increase the likelihood of providing constructive feedback that aligns with the model's learning objectives, such as successfully passing rationality tests. Additionally, intentionally integrating human biases and heuristics into AI systems holds potential benefits in specific scenarios. While humans often exhibit cognitive biases and irrational behavior, deliberately embedding these biases into AI systems can prove advantageous. Understanding the underlying reasons behind human biases, such as the efficiency of heuristics in decision-making, offers valuable insights into their potential value in augmenting AI systems. However, it is crucial to distinguish between unintentionally incorporating human biases due to training data and purposefully integrating biases and heuristics into AI systems. While unintentional biases can lead to adverse outcomes, the deliberate integration of heuristics can enhance decision-making, particularly in situations with time constraints or limited resources [20, 33].

4.2 Auditing and Transparency

Human feedback plays a pivotal role in constructing rational models as we endeavor to refine AI training methodologies. As societal scrutiny over responsible governance frameworks for AI systems intensifies, transparency becomes paramount in enhancing and evaluating this feedback. Transparency and auditing are indispensable mechanisms within the governance framework of Language Models (LLMs) and reinforcement learning with RLHF, ensuring accountability and aiding in risk mitigation, particularly in competitive pressures that may obscure crucial governmental mandates such as transparency. Therefore, there exists an urgent necessity for independent auditing, evaluations, certification, vigilant post-deployment monitoring, and control over critical resources such as hardware and data [6, 1, 24].

A thorough understanding of the model is crucial for improved auditing and transparency. This requires disclosing several vital elements, including the specific RLHF algorithm used, the composition and size of the dataset, the methods employed for human feedback collection, and any preprocessing or filtering steps applied to the feedback data. Additionally, providing details about the model architecture, hyperparameters, training duration, and performance metrics is essential.

Enhanced transparency in these areas cultivates trust and enables stakeholders to assess the model's robustness, fairness, and potential biases more effectively. Furthermore, it facilitates experiment replication, allowing for independent verification of results and fostering collaborative research endeavors to advance safer and more reliable AI systems [6]:

- Pretraining Process Description: Detailed information about the pretraining process, including data sources and potential biases, is crucial for grasping the model's development context.
- Selection and Training of Human Evaluators: Transparency in how human evaluators are chosen and trained ensures the integrity and reliability of the feedback provided.
- Process for Selecting Feedback Examples: Disclosing the method for selecting feedback examples and safeguarding against data poisoning attacks is essential for maintaining data integrity.

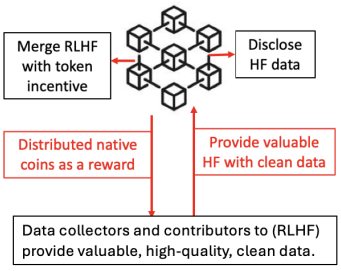

Figure 3: A possible scenario for utilizing blockchain to incentivize human feedback.

- Types of Human Feedback Used: Specifying the types of human feedback utilized offers insight into potential risks and the effectiveness of the feedback process.
- Quality Assurance Measures, such as feedback collection and inter-rater consistency, ensure the reliability and consistency of gathered feedback.

Disclosing details of human feedback offers stakeholders invaluable insights into the feedback process and its impact on model development and effectiveness. However, maintaining comprehensive records and ensuring privacy throughout training poses a significant challenge. Overcoming this obstacle necessitates collaborative efforts from stakeholders across sectors to integrate rigorous standards into AI governance frameworks. Integrating blockchain technology presents promising solutions to enhance transparency and auditing within AI systems [23]. Blockchain facilitates secure logging of the feedback loop, ensuring data immutability and transparency. This capability allows stakeholders to verify the authenticity and integrity of feedback records, enabling transparent and tamper-proof auditing of human feedback data. Furthermore, blockchain technology facilitates the evaluation of data providers by leveraging feedback mechanisms while safeguarding anonymity through innovative techniques like Zero-knowledge Proof [10]. This approach empowers individuals to verify credentials without compromising sensitive information. By promoting accountability and rewarding high-quality contributions [12], the integration of blockchain technology offers a compelling solution to enhance transparency and auditing in AI systems' feedback processes. In essence, the integration of blockchain technology holds considerable promise in addressing the inherent challenges of transparency and accountability in AI systems' feedback loops. However, despite its vast potential, further research is essential to fully explore and exploit its diverse applications. Figure 3 illustrates a possible scenario for utilizing blockchain to incentivize human feedback.

4.3 Future Directions

In pursuing enhanced data quality in human feedback and auditing processes, upcoming initiatives are poised to embrace innovative methodologies, particularly by incorporating decentralized autonomous organizations (DAOs) built on blockchain technology [2, 15, 34]. With the growing demand for high-quality datasets, the fusion of DAOs with token incentives emerges as a powerful strategy to bolster data reliability and inclusively. By leveraging token incentives, platforms can incentivize user participation in data collection, cultivating a broader and more engaged user community while enriching the learning experience. Simultaneously, DAO integration enables users to influence data quality standards directly, promoting transparent decision-making and community-driven improvements. Within the DAO framework, collaborative endeavors can establish and maintain robust data quality protocols, including validation mechanisms and strategies to mitigate bias. This synergistic relationship between token incentives and DAO governance motivates user contributions and safeguards the integrity of data standards. Ultimately, this collaborative approach holds significant potential for advancing research in Reinforcement Learning from Human Feedback (RLHF), fostering innovation, and ensuring the reliability and diversity of datasets essential for the progress of machine learning algorithms.

4.4 Limitations

Developing a large-language model-based RLFB with universal democratic norms presents a significant challenge due to the limitations of rational decision-making. Users often express preferences beyond rational considerations, complicating alignment with diverse intentions, particularly in artificial general intelligence (AGI). Designing RLFB models that respect individual inclinations within the rationality framework presents a formidable challenge in achieving universal alignment across AI Large Language Models (LLMs) [22]. Moreover, acknowledging human irrationality adds another layer of complexity to developing AI systems. Despite endeavors to create rational LLMs, human cognitive biases frequently lead to deviations from rational behavior, necessitating their integration into AI systems. Understanding human irrationality benefits psychology and enhances artificial intelligence, emphasizing the challenge of designing AI systems that emulate human-like decision-making while leveraging human biases [17, 11]. Conversely, assessing Natural Language Generation (NLG) systems via human ratings frequently needs to acknowledge the creative dimensions of human cognition. While aggregating ratings across annotators aims to encapsulate collective preferences, it often overlooks the subtleties of creativity. Despite continuous efforts to enhance NLG evaluation techniques, these constraints persist. Thus, there is a pressing need for the creation of inventive evaluation methodologies that not only acknowledge but also embrace human irrationality, recognizing its potential to spur creativity, for a more precise assessment of NLG systems [9].

5 CONCLUSION

Large Language Models (LLMs) have made significant progress in specific capabilities related to rational thinking and complex cognitive processes. This study utilized rationality tests to evaluate LLMs' performance, revealing challenges in answering simple questions accurately. Further exploration of LLMs' capabilities in natural language processing and narrative generation is needed. Understanding LLMs' rationality mechanisms and refining evaluation methodologies are crucial for addressing these challenges.

Collaboration across disciplines and innovative approaches are essential to fully harness LLMs' potential in reasoning and advancing artificial intelligence. Moreover, rationality and irrationality are fundamental in AI, necessitating a multidisciplinary perspective. Recognizing the potential benefits of irrational behavior in specific contexts prompts ongoing research into methods for effectively engaging with irrational agents. In human-AI interaction, accommodating human irrationality is vital, given humans' role in providing feedback. While cognitive biases may enhance artificial agent performance, system design must accommodate human irrationality. Exploring how an artificial agent's rationality influences human-AI interaction dynamics underscores the importance of addressing lingering questions as AI becomes more integrated into daily life.

REFERENCES

[1] Avinash Agarwal, Harsh Agarwal, and Nihaarika Agarwal. 2023. Fairness score and process standardization: framework for fairness certification in artificial intelligence systems. *AI and Ethics*, 3, 1, 267–279.

[2] Dana Alsagheer, Lei Xu, and Weidong Shi. 2023. Decentralized machine learning governance: overview, opportunities, and challenges. *IEEE Access*.

[3] Christopher Anderson and Sophia Drossopoulou. 2003. BabyJ: from object based to class based programming via types. *WOOD*, 82, 7, 53–81.

[4] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120, 6, e2218523120.

[5] Alexander P Burgoyne, Cody A Mashburn, Jason S Tsukahara, David Z Hambrick, and Randall W Engle. 2023. Understanding the relationship between rationality and intelligence: a latent-variable approach. *Thinking & Reasoning*, 29, 1, 1–42.

[6] Stephen Casper et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

[7] Nick Chater and Mike Oaksford. 2001. Human rationality and the psychology of reasoning: where do we go from here? *British Journal of Psychology*, 92, 1, 193–216.

[8] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

[9] Kawin Ethayarajh and Dan Jurafsky. 2022. The authenticity gap in human evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, (Eds.) Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, (Dec. 2022), 6056–6070. DOI: 10.18653/v1/2022.emnlp-main.406.

[10] Uriel Fiege, Amos Fiat, and Adi Shamir. 1987. Zero knowledge proofs of identity. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, 210–217.

[11] Gerd Gigerenzer and Daniel G Goldstein. 1996. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103, 4, 650.

[12] Fabian Glaesser. 2018. Does the transparent blockchain technology offer solutions to the algorithmic fairness problem? *Available at SSRN 3378071*.

[13] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

[14] Janmanchi Harika, Palavadi Baleeshwar, Kummari Navya, and Hariharan Shanmugasundaram. 2022. A review on artificial intelligence with deep human reasoning. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 81–84. DOI: 10.1109/ICAAIC53929.2022.9793310.

[15] Samer Hassan and Primavera De Filippi. 2021. Decentralized autonomous organization. *Internet Policy Review*, 10, 2, 1–10.

[16] Max Hendrix. 2023. Intricacies of agency: rational choice, behavioral economics, and our normative commitments.

[17] Daniel Kahneman, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

[18] Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2023. Chatgpt's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13, 1, 4569.

[19] Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. 2023. The history and risks of reinforcement learning and human feedback. *arXiv e-prints*, arXiv–2310.

[20] Olivia Macmillan-Scott and Mirco Musolesi. 2023. (ir) rationality in ai: state of the art, research challenges and open questions. *arXiv preprint arXiv:2311.17165*.

[21] Ken Manktelow. 2012. *Thinking and reasoning: An introduction to the psychology of reason, judgment and decision making*. Psychology Press.

[22] Abhilash Mishra. 2023. Ai alignment and social choice: fundamental limitations and policy implications. *arXiv preprint arXiv:2310.16048*.

[23] Ahmed Afif Monrat, Olov Schelén, and Karl Andersson. 2019. A survey of blockchain from the perspectives of applications, challenges, and opportunities. *IEEE Access*, 7, 117134–117151.

[24] Marçal Mora-Cantallops, Salvador Sánchez-Alonso, Elena García-Barriocanal, and Miguel-Angel Sicilia. 2021. Traceability for trustworthy ai: a review of models and tools. *Big Data and Cognitive Computing*, 5, 2, 20.

[25] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

[26] Megan K O'Brien and Alaa A Ahmed. 2016. Rationality in human movement. *Exercise and sport sciences reviews*, 44, 1, 20–28.

[27] Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.

[28] Stuart Russell. 2016. Rationality and intelligence: a brief update. *Fundamental issues of artificial intelligence*, 7–28.

[29] Sebastian Thrun and Michael L Littman. 2000. Reinforcement learning: an introduction. *AI Magazine*, 21, 1, 103–103.

[30] Amos Tversky and Daniel Kahneman. 1988. Rational choice and the framing of decisions. *Decision making: Descriptive, normative, and prescriptive interactions*, 167–192.

[31] Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: a survey. *arXiv preprint arXiv:2307.12966*.

[32] Peter Cathcart Wason and Philip Nicholas Johnson-Laird. 1972. *Psychology of reasoning: Structure and content*. Vol. 86. Harvard University Press.

[33] Ying Wen, Yaodong Yang, Rui Luo, and Jun Wang. 2019. Modelling bounded rationality in multi-agent interactions by generalized recursive reasoning. *arXiv preprint arXiv:1901.09216*.

[34] Andrej Zwitter and Jilles Hazenberg. 2020. Decentralized network governance: blockchain technology and the future of regulation. *Frontiers in Blockchain*, 3, 12.