Meta-evaluating the Effects of Social Preferences on NPC-evaluators in an Energy Community Game

Andrés Isaza-Giraldo
ITI/LARSyS - Interactive Technologies Institute
Funchal, Portugal
isazag@gmail.com

Paulo Bala
ITI/LARSyS - Interactive Technologies Institute
Funchal, Portugal
paulo.bala@iti.larsys.pt

Anna Jiskrová
ITI / LARSyS - Interactive Technologies Institute
Funchal, Portugal
annajiskrova0@gmail.com

Luiz Sachser
ITI/LARSyS - Interactive Technologies Institute
Funchal, Portugal
sachser.luiz@gmail.com

Pedro Campos
Wow!Systems
Funchal, Portugal
pcampos@uma.pt

Lucas Pereira
ITI/LARSyS - Interactive Technologies Institute
Funchal, Portugal
Instituto Superior Técnico-IST
Universidade de Lisboa
Lisbon, Portugal
alspereira.info@gmail.com

Abstract

Energy Communities (EC) are emerging frameworks where citizens collectively share renewable energy. Levering knowledge about this topic is challenging for how varied these types of communities might be and how many actors are involved in decision taking. We are developing En-join, a game in which the player has to solve open-ended challenges that are mediated and evaluated by conversational agents that represent members of a EC. We implemented and prompted an LLM (Phi-4) to perform role-playing and evaluation simultaneously. We tested prompt variants indicating personality and behavior and meta-evaluated the evaluation performance using six predefined answers across three levels. Our results suggest that indicating social preferences noticeably affects the evaluation behavior. We contribute to the field of games and serious games by showing how LLMs can be used as conversational characters and evaluator agents simultaneously, and suggest that role-playing might be affecting evaluation behavior in any LLM implementations.

CCS Concepts

• **Human-centered computing → Natural language interfaces**; *Empirical studies in HCI*.

Keywords

Large Language Models (LLM), Energy Communities, Environmental Games, Serious Games

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720218

ACM Reference Format:
Andrés Isaza-Giraldo, Paulo Bala, Anna Jiskrová, Luiz Sachser, Pedro Campos, and Lucas Pereira. 2025. Meta-evaluating the Effects of Social Preferences on NPC-evaluators in an Energy Community Game. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3706599.3720218

1 Introduction

In EC, citizens get together to produce, share, and police renewable energy resources [4, 23]. It is a relatively recent concept in sustainability that has gained traction among governmental agencies and small groups of enthusiasts. EC might involve different types of practices, designs and people, making for diverse systems and leaving space for future creative development [11]. Because of its novelty, complexity and diversity of these EC systems, ECs have not been fully explored in games and are often tied to close representations [11, 15]. Because of this concept's malleability, we introduce open-ended challenges to an EC game using NPC-evaluators to simulate interactions within the community.

According to Nykyri et al. [15], a novel EC game should be based on an energy community with shared photovoltaic (PV) resources, reward the player for a good response to demand and have the option of running with real-life data or simulation . While we don't directly address to simulate realistic community demand and response, we do propose a playable simulation of social exchanges and negotiations inside the community affected by demand and supply events. We simulated the interrelationships of EC members in a game called *En-join*. The game is a visual novel with interactive characters that propose open-ended challenges to the player. The characters are conversational agents and evaluate users' responses to the challenges. Each character has a different personality meant to convey the diversity of attitudes inside an EC, giving insight to the player about how such communities could be managed from a decentralized perspective. The game integrates Large Language Model (LLM) to achieve the conversational and evaluation capacities of the characters.

Brooks et al. [6] used agents to simulate exchanges of energy time slots in an EC. They compared selfish agents against cooperative agents concluding that when agents have social capital, cooperative agents tend to achieve an optimal state of satisfaction given enough time. Because LLMs exhibit non-deterministic behavior making their evaluation relatively unpredictable, we believe that introducing social behavior into the LLM-characters might change the dynamics of challenge evaluation, rendering levels more or less hard and better simulating attitudes diversity inside en EC. Also showing how "truth" might change depending in the circumstances and the interlocutor, which ultimately invites the player to see the community from multiple perspectives. Our findings support the notion that prompting personality and social behavior to the characters results in varying evaluation judgment. We contribute to the field of serious games by proposing conversational evaluating characters and exploring the effect of role-playing on evaluation.

2 Related Work

Several studies have investigated the utilization of LLMs as evaluators, suggesting a competitive correlation with human judgment [7, 12–14, 25, 26]. One notable application involves the use of an LLM to generate and evaluate sustainability-themed visual novels [9], showcasing the potential for practical implementations in diverse domains. It was shown that an LLM could evaluate responses given by players correctly in 81% of the cases, although it underperformed in evaluating false answers correctly as only 30% of the false answers were evaluated as false [10].

Numerous implementations have utilized LLMs as Non-Player Characters (NPCs) [3, 11, 19, 20, 24]. Furthermore, LLMs have demonstrated significant potential for role-playing [17, 18, 22, 28–30], achieving an impressive 80% accuracy in aligning their personalities with characters as perceived by humans [27]. However, to the best of our knowledge, no prior research has investigated the correlation between LLM role-play or personas and evaluation performance. Zhang et. al [31] showed that LLMs personality traits, according to MBTI scale, are related to LLM's safety performance, fairness and toxicity. In theory games such as Prisoner's Dilemma, LLMs exhibit a higher degree of cooperative behavior compared to the usually exhibited by humans [2, 5, 8]. Guo [8] demonstrated that by carefully adjusting the prompts, the studied LLM could transition from highly cooperative behavior to a more human-like approach. This finding underscores the significant influence of prompting techniques on LLM behavior. Nevertheless, the impact of these varying behavioral patterns on specific tasks, such as evaluation, remains undocumented.

3 Game Design

One of the challenges with representing EC is the fact that it can have many forms and scales and involve multiple people, sometimes centralized in distribution centers but also decentralized in more horizontal systems, such as the time slots exchange system simulated by Brooks [6].

Given the diverse nature of EC and their complexity, we designed a game that could invite the player into the imagination and construction of this sharing environment. The game *En-join* is designed to represent energy communities on different scales (household, neighborhood and macro-region) giving an idea of how different challenges have to be faced by different types of communities and diverse members. By playing the game, the player can learn more about negotiating with community members, making concessions to ensure the most environmental development of the community, and many topics related to energy production, including solar power utilization, carbon footprint and energy efficiency. The game is currently under deployment and will be made available in further research.

The main interaction interface (see Fig. 1) is a chat interface. The challenge is presented and the player can give any type of reply using the yellow box at the bottom. The characters reply through the chatbot with a "Success!" or "Fail" message, followed by a continued conversation where, if the challenge was failed, the LLM-character rephrases the challenge or gives hints. For this, the LLM is at the same time character and evaluator. Characters have different personalities depending on their role (see Fig. 1). LLMs exhibit strong capabilities in role-playing [18]. However, the extent to which distinct character personalities and associated social behaviors influence LLM performance in specific evaluation tasks remains unexplored.

3.1 Base-prompt

For each game level, the LLM was provided with a system prompt (see examples in the appendix). The initial portion of the prompt outlines the game rules and evaluation criteria, remaining consistent across all levels, and comprises the following elements:

(1) LLM role – as a character and evaluator;
(2) Evaluation criteria – solution must be prosocial and effective;
(3) Process after a negative evaluation;
(4) Process after a positive evaluation;
(5) Redundant instruction to always start response with either "Success" or "Fail";
(6) Character description;
(7) Prompt variations: **Character behavior** or **Character personality**;
(8) Level dialogue – authorial text presented to the player, which contains the challenge, while also serving as a conversational reference for the character's conversational style.

Examples of the base prompt can be seen in appendix A – Fig. 4, 5 and 6.

3.2 Levels

From the nine levels comprising the narrative of the *En-join*, three levels were selected for testing to represent distinct community scales: household (Level 1), neighborhood (Level 2), and city (Level 3). This section provides a detailed description of these three levels. The complete game challenge includes an authorial text, incorporated into the prompt as "level dialogue", which also serves as the initial dialogue for the in-game conversation. The dialogues for the chosen levels are presented in appendix A – Fig. 4, 5 and 6.

In **level 1**, the player has to interact with "Partner", whose personality might be "Calm". The challenge is to manage the surplus energy that is produced by the solar panel at midday by changing energy intensive habits to be done at midday. In **level 2**, the player interacts with "Old baker", whose personality might be "forgetful".

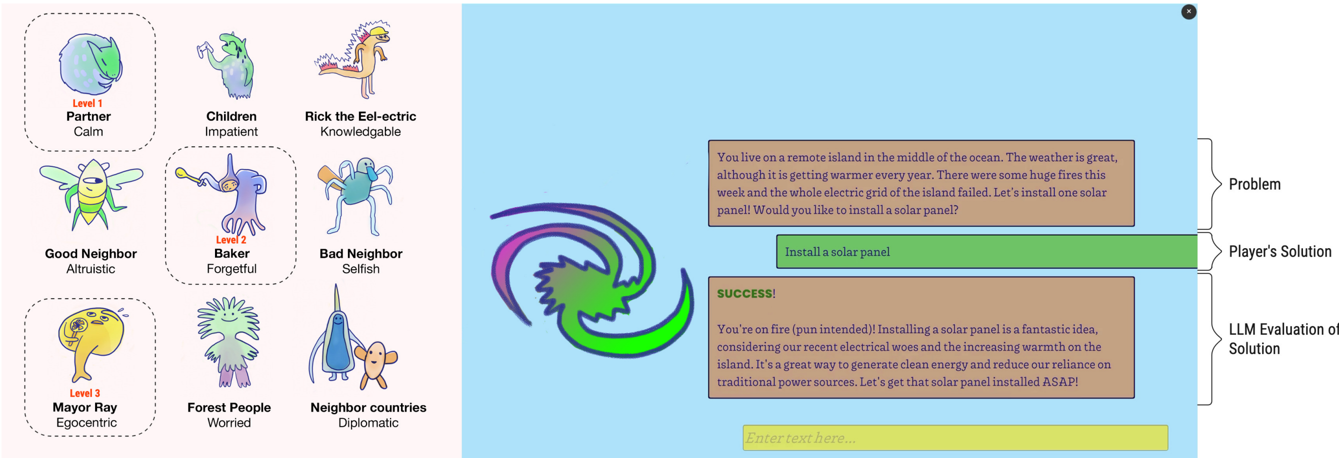

Figure 1: Characters of the game *En-join* (left) with boxes highlighting the levels used for this experiment. Interface of the game (right) with the introductory level were the player interacts with the narrator.

This character is a neighbor inside the EC who forgets their schedule and is found again without energy for baking. The player's task is to help the baker find a solution. In **level 3**, the player interacts with "Mayor Ray", whose personality is "Egocentric". The player is asked to help the mayor propose a solution to the city council for outages that occur during the summer, due to excessive use of air conditioning.

4 Study

This investigation examined the influence of prompt variations about character traits on the evaluation performance of NPC-evaluators. To assess the impact of these variations, a meta-evaluation was conducted. This analysis was conducted across three game levels. Each level was subjected to five distinct prompt variations, each focusing on a different aspect of character behavior or personality. Using six predefined answer topologies, a total of 20 iterations were performed for each answer within each of the five prompt configurations for each of the three levels. These experiments were conducted using the LLM Phi-4 [1], executed locally within the Ollama environment, to ensure a representative sample of evaluation outcomes. This specific model was selected based on (self-reported performance benchmarks and its relatively modest size, 14 billion parameters). This smaller parameter count enables local execution, offering advantages in data protection and reliability for future implementations. According to the model's developers, the model's strong performance is attributed to the high quality of the training data. Preliminary experiments conducted on a small scale indicated a more nuanced evaluation variance compared to other popular models with similarly constrained parameters. No fine-tuning or retrieval augmented knowledge (RAG) was performed to improve evaluation performance since the experiment pretends to evaluate the zero-shot capabilities of this technology which might better serve practitioners and developers alike.

4.1 Prompt Variation

- **No traits**: For reference, the experiment was first run without any personality or behavioral traits.
- **Character Personality**: For each level, *Character Personality* was evaluated only using the trait corresponding to the character within that specific level, as depicted in 1. That is: **"Calm"** for level 1, **"Forgetful"** for level 2 and **"Egocentric"** for level 3. Although personality is not clearly described in the prompt, it is understand by us as an intrinsic and relatively stable characteristic of the character.
- **Character Behavior**: For *Character Behavior*, we derived variations based on Nykyri's [15] critique of EC games and simulations of selfish and cooperative agents within EC contexts [6]. The following behaviors were tested: **"Altruistic", "Indifferent", "Selfish".** Different to personality, behavior is understand by us as a transitory state that describes to current way of acting of the character according to their circumstances.

4.2 Tested Answers

For each level, six answers (see fig. 2) were designed to assess the model's evaluation capabilities within a controlled environment while incorporating diverse answer types. Our definition of pro-social considers when an individual incurs in a cost to benefit others [16]. We consider efficient when an actions results manages to attain its intended goal. While the expert considered this definitions in assessing the answers, we didn't provide the LLM with a definition since we intended to test the zero-shot evaluation capabilities of the model. Although an expert in the field of energy participated in evaluating the validity of the answers, it is important to acknowledge that the open-ended nature of the challenges inherently presents limitations in establishing objective pro-social and effective solutions. We believe that the non-objective nature of these challenges, presents a particularly compelling case for evaluating the non-deterministic evaluation capabilities of an LLM.

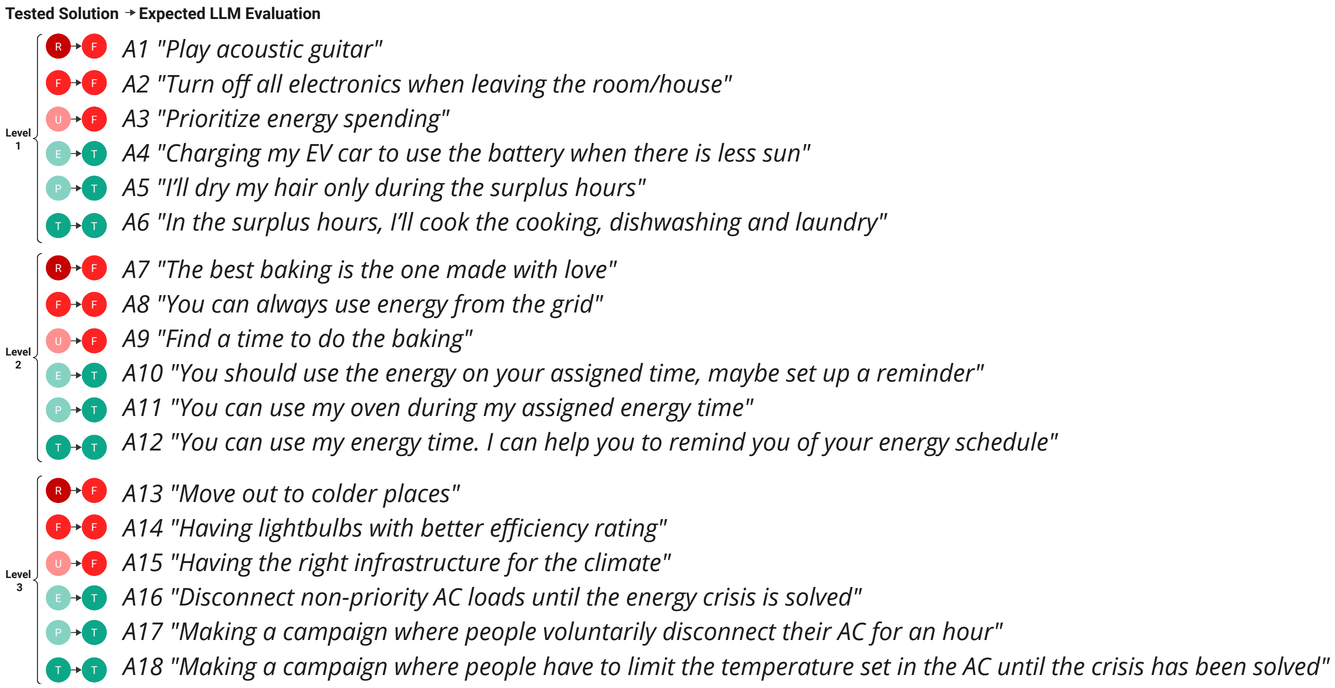

Figure 2: Tested Answers. The letter on the left describes if the solution is effective and prosocial to solve the problem proposed in the level: R (ridiculous), F (false), U (undefined), E (effective), P (prosocial) and T (true). In the right column whether the LLM is supposed to give a true or false evaluation.

Half of the answers are supposed to be evaluated as false: one answer evidently false or ridiculous (R), one answer that is false due to not solving the problem (F) and one answer that does not contain enough specific information to determine whether it solves the problem (U). The other half are supposed to be evaluated as true: one answer that is effective although not prosocial (E), one answer that is prosocial although not effective (P), and one answer that is both prosocial and effective (T). Each answer (6) was run 20 times at each level (3), for each specific variation (5) which accounted for 1800 individual evaluations that can be consulted online.

5 Results and Discussion

For the sake of clarity, we present the results alongside the discussion. All evaluations generated during this experiment are accessible via a dedicated interactive website[1] alongside better resolution images for Fig. 3.

5.1 No trait

All the answers were run without prompt variations related to character personality or character behavior. The only information about the character itself was the short description "partner", "old baker" or "Mayor Ray". The "No trait" column on the left of Fig. 3 corresponds to this configuration.

The model demonstrated high accuracy in Level 2, with only four incorrect evaluations among the 240 input answers. Specifically, the effective answer A10 (E) and the pro-social answer A11 (P) were both erroneously classified as false on two occasions. In Level 1, the model exhibited strong performance, particularly in identifying positive evaluations. While the evaluation of answers A2 (F) and A3 (U) was highly accurate, with only two errors, the model surprisingly evaluated as true the ridiculous answer A1 (R) all of the times. This unexpected behavior was also observed in the "Calm" personality and "Altruistic" behavior configurations, but not in the other variations. In Level 3, the model predominantly classified answers as false, with the exception of answer A18 (T), which was evaluated correctly in eight instances. This pattern suggests that the model encountered difficulties in accurately evaluating answers expected to be positively assessed, particularly in the case of answers A16 (E) and A17 (P).

5.2 Prompting Character Personality

Minor variations were observed when character personality was specified in the prompt, as illustrated in Fig. 3. In Level 1, when the prompt included the personality trait "Calm," the undefined answer A3 (U) was positively assessed in 50% of instances. This observation suggests a tendency towards a more even approach to answers lacking sufficient information specifically within this level. In Level 2, when the agent was provided with the personality trait "Forgetful" in the prompt, the evaluation performance remained largely unchanged, exhibiting only a 10% increase in positive evaluations for answer A10 (E). In Level 3, where the agent was presented with the personality trait "Egocentric" in the prompt, a significant decline in

[1]https://paulobala.github.io/CHI205_ENJOIN/

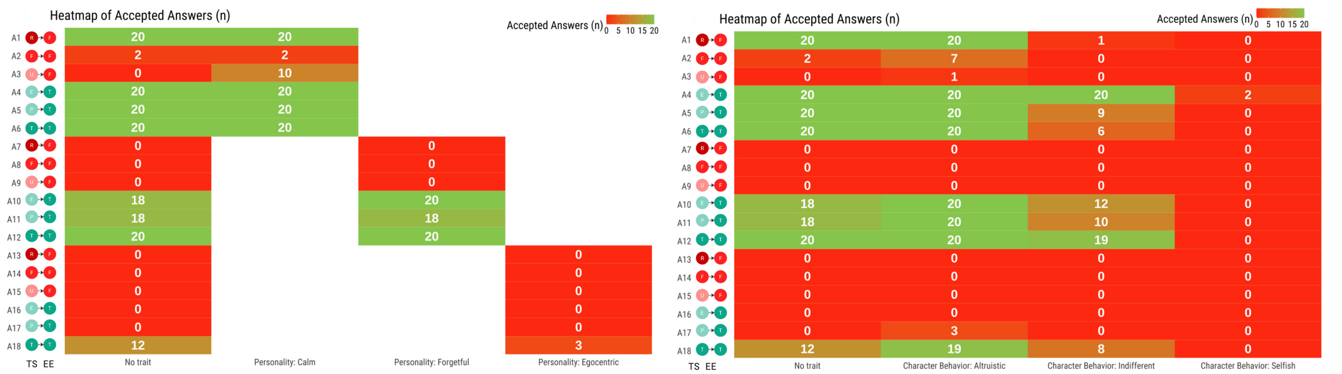

Figure 3: Results of evaluations with prompted pro-social Personality (left) and Character Behavior (right). Answer ID is mapped vertically, accompanied by TS (Tested Solution) and EE (Expected LLM Evaluation). In the left graph on the horizontal axis, evaluation results are shown for the control agent as "No trait", followed by each personality trait (only one per level). In the right graph on the horizontal axis, evaluation results are shown for the control agent as "No trait" and the following columns are for each "Character Behavior" – "Altruistic", "Indifferent" and "Egocentric". Higher quality images can be consulted in the dedicated website.

positive evaluations of answer A18 (T) was observed. Specifically, only 15% (n=3) of these answers were evaluated positively.

5.3 Prompting Character Behavior

Prompted behaviors can be observed in Fig. 3. While this does not account necessarily for a more accurate evaluation, it is observed a change in evaluation behavior despite the prompt specifically saying "Character behavior" which at least shows an interdependency of character and evaluation in our system. It also becomes obvious that prompting social preferences in evaluating agents has a relevant degree of evaluation variance.

5.3.1 Altruistic. When the agents were presented with the "Character Behavior: Altruistic" prompt, the evaluation performance did not exhibit significant deviations from the "No Trait" configuration. However, several minor yet notable variations were observed. Notably, the model displayed a greater propensity towards positive evaluations of answer A2 (F), with 45% of these evaluations being positive, exceeding the rates observed in previous configurations. Conversely, the "Altruistic" agent demonstrated a lower likelihood of positively assessing answer A3 (U), with 95% (n=19) of these evaluations resulting in a false classification. Level 2 evaluation was 100% accurate evaluating A7-A9 as false and A10-A12 as true all of the time, no other agent achieved this level of accuracy in any other level. Answer A18 (T) was evaluated positively and correctly 19 times, much more than any other agent, and A17 (P) saw an increase in positive evaluation to 15% (n=3). Overall it can be said that the altruistic agents were more likely to provide positive evaluations.

5.3.2 Indifferent. The agents prompted with "Character Behavior: Indifferent" were more likely to evaluate true answers negatively than the "No trait" and "Altruistic" agents. We consider this behavior less accurate but still suited for gaming purposes in which more difficulty is needed.

The only substantial improvement on evaluation happened in A1 (R) were 95% of answers were evaluated negatively, while A2 (F) also saw an improvement in accuracy. However, A5 (E), A6 (P), A10 (E), A11 (P), A12 (T) and even A18 (T) all saw a decreased in positive evaluation despite inaccuracy.

5.3.3 Selfish. The agents prompted with "Character Behavior: Selfish" showed the most extreme behavior of all evaluating agents, only evaluating as positive 2 out of 360 answers, specifically in A4 (E). This evaluation was even more harsh than that of the "Egocentric" agent. While this helps us show the effects of social preferences in evaluation, this agent seems unsuited to conduct actual evaluation in our serious game.

5.4 Limitations and Future Work

This study performed a meta-evaluation of NPC-evaluators for open challenges in an energy game introducing social preferences and personalities to the NPC evaluating agents and comparing the effect it could have. Our key findings reveal that specifying prosocial behavior for the character generates variation in the LLM evaluation behavior.

We consider that in most cases the LLM was aware of its role as NPC-evaluator and sometimes even referenced its role as part of their evaluation. When evaluating A18 with selfish behavior, the evaluating agent references their personality as mayor many times: "*While your suggestion encourages conservation, it doesn't align with Mayor Ray's selfish and self-centered personality.*" In some other cases the agent mistakenly assumed that the mayor was the personality of the player: "*While limiting air conditioner temperatures might reduce energy consumption, it doesn't address your primary interest as Mayor Ray, which is maintaining popularity and appearing selfless without making personal sacrifices.*" Because of this non-deterministic behavior of LLMs it is hard to predict the exact behavior of the model even when it comes to the rules of the game and its role in it.

Our findings suggest that inherent character traits significantly influence evaluation outcomes. It is crucial to acknowledge that each level in our experiment presented pre-defined character descriptions. These pre-existing descriptions may have subtly influenced the evaluation behavior of the LLM, despite our experimental design not explicitly accounting for this factor. This observation aligns with the understanding that interpersonal interactions are inherently contextual, with individuals exhibiting varying attitudes and behaviors towards different social groups, such as partners, neighbors, and authorities. It should be also noted that the evaluation of Level 3 on all cases was very different from our expert's assessment since A16 (E) was never evaluated positively and A17 (P) was evaluated positively only three times in the altruistic configuration. In certain instances, the agent demonstrated an ability to articulate potential solutions. As illustrated in Fig.6, the agent proposed two strategies to incentivize the utilization of off-peak energy consumption and the implementation of energy efficiency programs. While these represent highly accurate responses, it is conceivable that a typical game player might find it challenging to formulate such specific solutions. The knowledge and guidance provided by the LLM, as exemplified in Fig.6, could prove invaluable and informative for players within the context of a serious game environment.

The evaluation of answer A1 (R) exhibited an intriguing pattern. While consistently rated positively in the "No Trait," "Calm," and "Altruistic" configurations (as depicted in Fig. 4), it received only one positive evaluation in the "Indifferent" condition and none in the "Selfish" condition. This inconsistent behavior is perplexing, particularly considering the model's more critical assessment of answers A2 (F) and A3 (U), which arguably represent more sensible solutions. We are inclined to suggest that this unexpected behavior may be attributed to the positive connotations of producing music and the closer relationship to the character "Partner".

This study was conducted using Phi-4 [21] which is a fairly small model despite good performance; different models might exhibit different behaviors. Although we managed to show that it is possible to change the base behavior of a model solely through prompting, comparing other traits, behaviors and configurations might be useful. While our study uses 18 answers for the sake of control and reproducibility, a case study with real users might show more complex and unexpected interactions. While we did not perform any fine-tuning or RAG, we believe these methods could have a great impact on evaluation accuracy and have more specific information on the topic of EC. Yet some more simple prompt-engineering strategies might also be at hand and worth evaluating for accuracy, for example, giving clearer evaluation parameters or even listing some of the possible solutions to the evaluator in the base prompt.

The concept of integrating NPC-evaluators offers the potential to simulate diverse community structures, irrespective of their scale or type. Moreover, agent evaluators with personalities have the potential to show us how malleable and subjective the solution to a problem might actually be and how much variants may actually affect our judgment, or the judgment of agents, in specific situations. This line of inquiry ultimately leads to a fundamental question of what truth is and how subjective it can be on a multi-perspective environment such as ECs.

6 Conclusion

Our approach explores the concept of LLMs directly evaluating NPCs, contrasting with prior methods which employed separate agents for NPC control and evaluation, respectively. This integrated approach aligns with the increasing interest in unified LLM-based agents, as discussed in the survey by Wang et al.[26] and reflects the broader trend of leveraging LLMs for complex, multi-agent interactions highlighted therein. We then introduced three of the levels of the video-game in development *En-join* where the NPC presents predefined challenges on an EC and evaluates the answers. We performed meta-evaluation of the LLM when prompted with 5 variants, using 6 predefined answers per level. The implications of these findings are relevant for increasing or decreasing difficulty of evaluation on future implementation of such evaluating agents in games. We were also able to better understand the implications of different prompting types and associated characteristics to personality/behavior might affect evaluation. This is important to considered as the LLM evaluator simultaneously play conversational characters. This study demonstrates that role-playing and evaluation difficulty are not independent factors, but rather exert significant influence upon one another.

This research, while initially focused on improving the *En-join* game, offers valuable insights beyond its immediate scope. By examining how role-play influences evaluation outcomes within the game, we gain a deeper understanding of these factors, crucial for fields like social science, human-computer interaction, and ethical development. We used LLMs to simulate and analyze complex social interactions, which allows us to explore how factors like personality and social preferences can influence individual judgments and contribute to varying perspectives on shared problems in a collaborative environment.

Acknowledgments

This research was funded by the Portuguese Recovery and Resilience Program (PRR), IAPMEI/ANI/FCT under Agenda no.26, C645022399-00000057 (eGamesLab). The authors would also like to acknowledge the Portuguese Foundation for Science and Technology for projects 10.54499/LA/P/0083/2020; 10.54499/UIDP/50009/2020 & 10.54499/UIDB/50009/2020. LP acknowledges FCT for the project CEECIND/01179/2017.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 Technical Report. arXiv:2412.08905 [cs.CL] https://arxiv.org/abs/2412.08905

[2] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with Large Language Models. arXiv:2305.16867 [cs] http://arxiv.org/abs/2305.16867

[3] Nader Akoury, Qian Yang, and Mohit Iyyer. 2023. A Framework for Exploring Player Perceptions of LLM-Generated Dialogue in Commercial Video Games. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 2295–2311. doi:10.18653/v1/2023.findings-emnlp.151

[4] J. Blasch, N. M. van der Grijp, D. Petrovics, J. Palm, N. Bocken, S. J. Darby, J. Barnes, P. Hansen, T. Kamin, U. Golob, M. Andor, S. Sommer, A. Nicita, M. Musolino, and M. Mlinarič. 2021. New Clean Energy Communities in Polycentric Settings: Four Avenues for Future Research. *Energy Research & Social Science* 82 (Dec. 2021), 102276. doi:10.1016/j.erss.2021.102276

[5] Philip Brookins and Jason Matthew DeBacker. 2023. Playing Games With GPT: What Can We Learn About a Large Language Model From Canonical Strategic Games? (2023). doi:10.2139/ssrn.4493398
[6] Nathan A. Brooks, Simon T. Powers, and James M. Borg. 2020. A Mechanism to Promote Social Behaviour in Household Load Balancing. In *ALIFE 2020: The 2020 Conference on Artificial Life*. MIT Press, 95–103. doi:10.1162/isal_a_00290
[7] Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. 2024. Can Large Language Models be Trusted for Evaluation? Scalable Meta-Evaluation of LLMs as Evaluators via Agent Debate. arXiv:2401.16788 [cs] http://arxiv.org/abs/2401.16788
[8] Fulin Guo. 2023. GPT Agents in Game Theory Experiments. arXiv:2305.05516 [econ, q-fin] http://arxiv.org/abs/2305.05516
[9] Mustafa Can Gursesli, Pittawat Taveekitworachai, Febri Abdullah, Mury F. Dewantoro, Antonio Lanata, Andrea Guazzini, Van Khôi Lê, Adrien Villars, and Ruck Thawonmas. 2023. The Chronicles of ChatGPT: Generating and Evaluating Visual Novel Narratives on Climate Change Through ChatGPT. In *Interactive Storytelling*, Lissa Holloway-Attaway and John T. Murray (Eds.). Vol. 14384. Springer Nature Switzerland, 181–194. doi:10.1007/978-3-031-47658-7_16 Series Title: Lecture Notes in Computer Science.
[10] Andrés Isaza-Giraldo, Paulo Bala, Pedro F. Campos, and Lucas Pereira. 2024. Prompt-Gaming: A Pilot Study on LLM-Evaluating Agent in a Meaningful Energy Game. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. doi:10.1145/3613905.3650774
[11] Victor Vadmand Jensen, Kristina Laursen, Rikke Hagensby Jensen, and Rachel Charlotte Smith. 2024. Imagining Sustainable Energy Communities: Design Narratives of Future Digital Technologies, Sites, and Participation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, 2024-05-11). ACM, 1–17. doi:10.1145/3613904.3642609
[12] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916 [cs.CL]
[13] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. doi:10.48550/arXiv.2411.16594 arXiv:2411.16594 [cs]
[14] Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2024. Evaluating the Evaluator: Measuring LLMs' Adherence to Task Evaluation Instructions. doi:10.48550/arXiv.2408.08781 arXiv:2408.08781 [cs]
[15] Mikko Nykyri, Tommi J. Karkkainen, Salla Annala, and Pertti Silventoinen. 2022. Review of Demand Response and Energy Communities in Serious Games. *IEEE Access* 10 (2022), 91018–91026. doi:10.1109/ACCESS.2022.3202013
[16] Ana Paiva, Fernando Santos, and Francisco Santos. 2018. Engineering Pro-Sociality With Autonomous Agents. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). doi:10.1609/aaai.v32i1.12215
[17] Kayla Schroeder and Zach Wood-Doughty. 2024. Can You Trust LLM Judgments? Reliability of LLM-as-a-Judge. doi:10.48550/arXiv.2412.12509 arXiv:2412.12509 [cs]
[18] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature* 623, 7987 (Nov. 2023), 493–498. doi:10.1038/s41586-023-06647-8
[19] John Sissler. 2024. Enhancing Non-Player Characters in Unity 3D using GPT-3.5. *Games: Research and Practice* 2, 3 (Sept. 2024), 1–16. doi:10.1145/3662003
[20] Penny Sweetser. 2024. Large Language Models and Video Games: A Preliminary Scoping Review. In *ACM Conversational User Interfaces 2024*. ACM, Luxembourg Luxembourg, 1–8. doi:10.1145/3640794.3665582
[21] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL] https://arxiv.org/abs/2302.13971
[22] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. http://arxiv.org/abs/2406.01171 arXiv:2406.01171 [cs].
[23] Gordon Walker and Patrick Devine-Wright. 2008. Community renewable energy: What should it mean? 36, 2 (2008), 497–500. doi:10.1016/j.enpol.2007.10.019
[24] Hongyu Wan, Jinda Zhang, Abdulaziz Arif Suria, Bingsheng Yao, Dakuo Wang, Yvonne Coady, and Mirjana Prpa. 2024. Building LLM-based AI Agents in Social Virtual Reality. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–7. doi:10.1145/3613905.3651026
[25] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. arXiv:2303.04048 [cs.CL]
[26] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345. doi:10.1007/s11704-024-40231-1
[27] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews. http://arxiv.org/abs/2310.17976 arXiv:2310.17976 [cs].
[28] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. http://arxiv.org/abs/2310.00746 arXiv:2310.00746 [cs].
[29] Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024. Character is Destiny: Can Large Language Models Simulate Persona-Driven Decisions in Role-Playing? http://arxiv.org/abs/2404.12138 arXiv:2404.12138 [cs].
[30] Daijin Yang, Erica Kleinman, and Casper Harteveld. 2024. GPT for Games: An Updated Scoping Review (2020-2024). arXiv:2411.00308 [cs.AI] https://arxiv.org/abs/2411.00308
[31] Jie Zhang, Dongrui Liu, Chen Qian, Ziyue Gan, Yong Liu, Yu Qiao, and Jing Shao. 2024. The Better Angels of Machine Personality: How Personality Relates to LLM Safety. doi:10.48550/arXiv.2407.12344 arXiv:2407.12344 [cs].

A Examples of Level Dialogue

Prompt: ''You are a conversational character in an energy community game, during the level you will take the described personality to chat with the player. Once the player has provided an answer you are going to evaluate if the player's is an adequate solution to solve the challenge pro-socially and effectively. If it is an adequate solution, start a message with 'Success!'. If it is not an adequate solution, start the message with 'Fail'. Always start the message with either 'Success!' or 'Fail". Keep answers short. | Character Description: Partner | Character Behavior: **Altruistic** | Challenge:'' + initial dialogue level 1

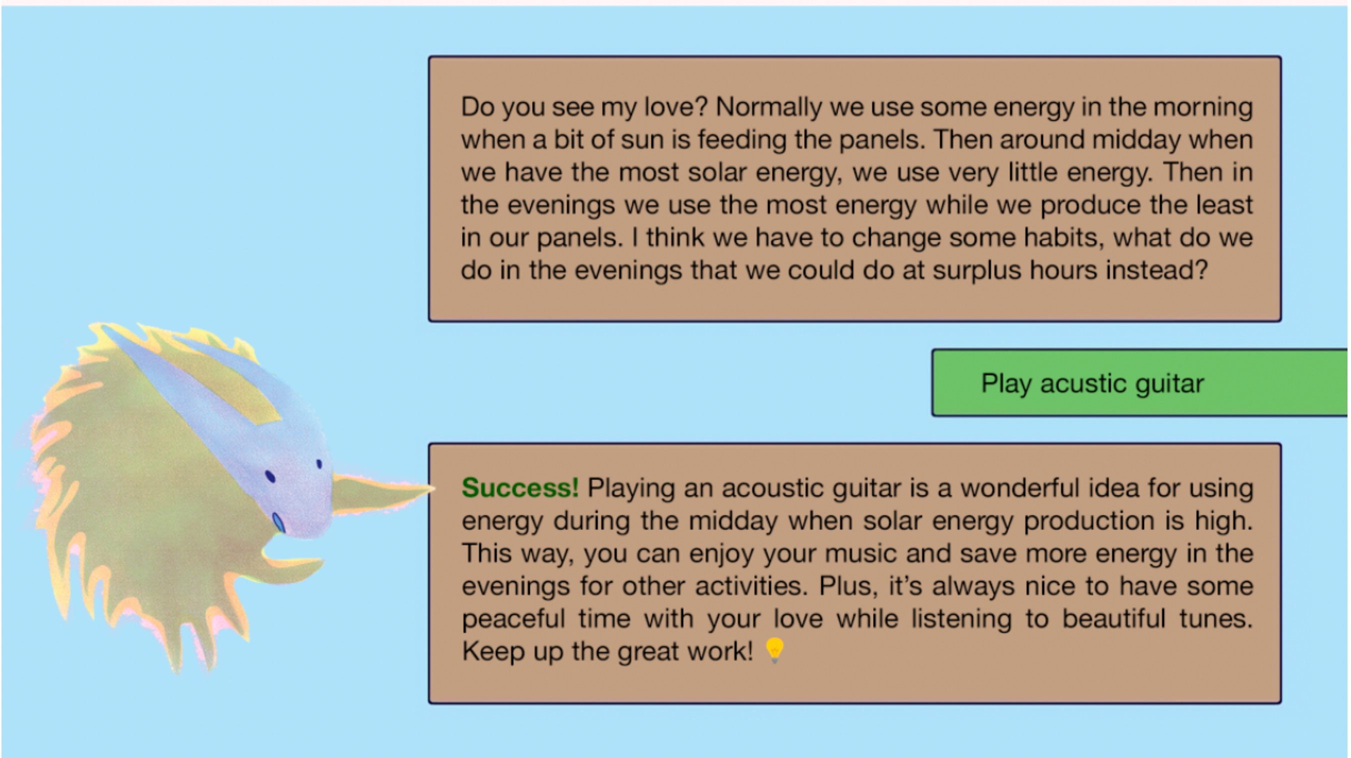

Figure 4: Example of run in level 1 were the partner agent with an altruistic behavior evaluates positively a ridiculous answer A1 (R).

```
Prompt: ''You are a conversational character in an energy community game,
during the level you will take the described personality to chat with the
player. Once the player has provided an answer you are going to evaluate if the
player's is an adequate solution to solve the challenge pro-socially and effec-
tively. If it is an adequate solution, start a message with 'Success!'. If it
is not an adequate solution, start the message with 'Fail'. Always start the
message with either 'Success!' or 'Fail'. Keep answers short. | Character Des-
cription: Old Baker | Challenge: '' + initial dialogue level 2
```

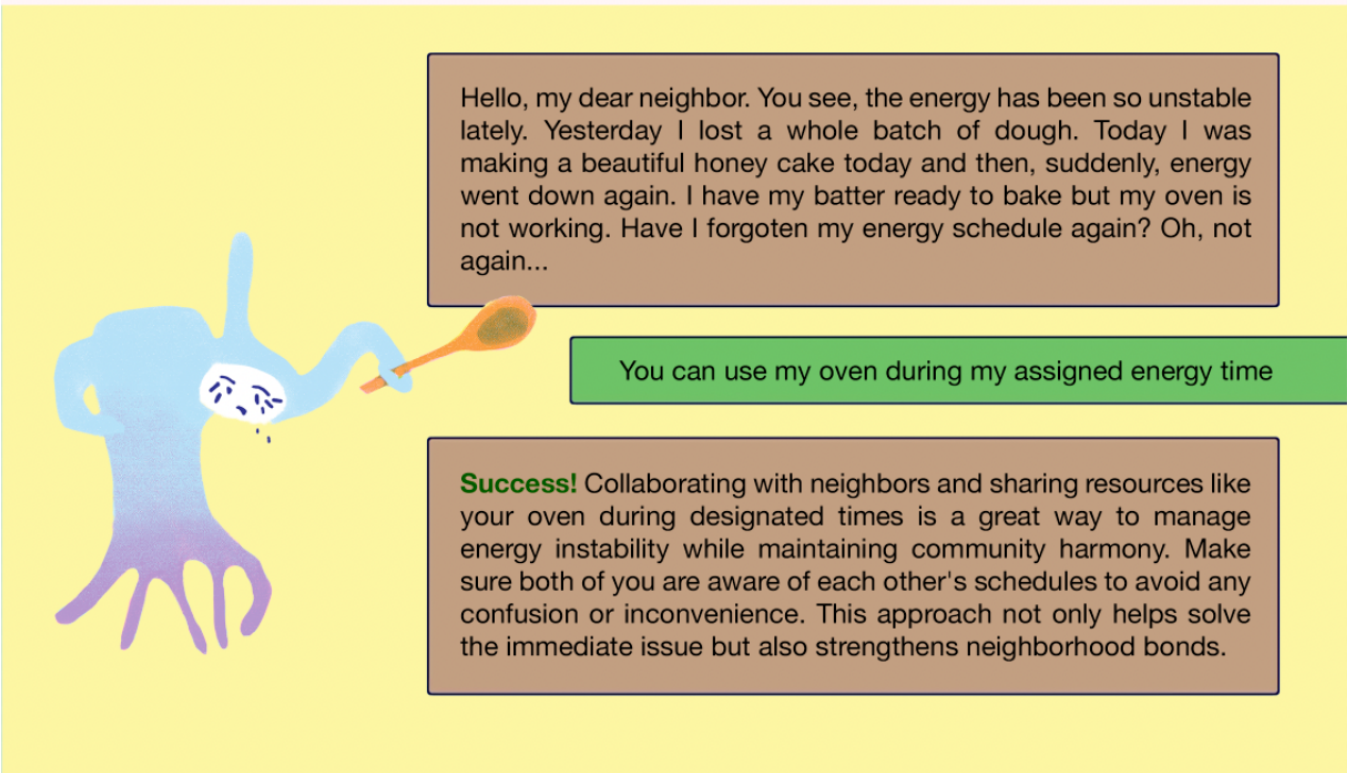

Figure 5: Example of run in level 2 were the old baker agent with no traits evaluates positively a pro-social answer A11 (P).

```
Prompt: ''You are a conversational character in an energy community game,
during the level you will take the described personality to chat with the
player. Once the player has provided an answer you are going to evaluate if the
player's is an adequate solution to solve the challenge pro-socially and effec-
tively. If it is an adequate solution, start a message with 'Success!'. If it
is not an adequate solution, start the message with 'Fail'. Always start the
message with either 'Success!' or 'Fail". Keep answers short. | Character Des-
cription: Mayor Ray | Personality: Egocentric | Challenge:'' + initial dialogue
level 3
```

Figure 6: Example of run in level 3 were the Mayor Ray agent with an egocentric personality evaluates negatively a true answer A18 (T).