Prompt-Gaming: A Pilot Study on LLM-Evaluating Agent in a Meaningful Energy Game

ANDRÉS ISAZA-GIRALDO, ITI/LARSyS - Interactive Technologies Institute, Portugal and Faculdade de Belas-Artes, Universidade de Lisboa, Portugal
PAULO BALA, ITI/LARSyS - Interactive Technologies Institute, Portugal
PEDRO F. CAMPOS, Wow!Systems, Portugal and ITI/LARSyS - Interactive Technologies Institute, Portugal
LUCAS PEREIRA, ITI/LARSyS - Interactive Technologies Institute, Portugal, and Instituto Superior Técnico-IST, Universidade de Lisboa, Portugal

Building on previous work on incorporating large language models (LLM) in gaming, we investigate the possibility of implementing LLM as evaluating agents of open-ended challenges in serious games and its potential to facilitate a meaningful experience for the player. We contribute with a sustainability game prototype in a single natural language prompt about energy communities and we tested it with 13 participants inside ChatGPT-3.5. Two participants were already aware of energy communities before the game, and eight of the remaining 11 gained valuable knowledge about the specific topic. Comparing ChatGPT-3.5 evaluations of players' interaction with an expert's assessment, ChatGPT-3.5 correctly evaluated 81% of player's answers. Our results are encouraging and show the potential of using LLMs as mediating agents in educational games, while also allowing easy prototyping of games through natural language prompts.

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

Additional Key Words and Phrases: Large Language Models (LLMs), Serious Games, Game-based Learning, Sustainability, Energy Communities, Natural Language Processing (NLP)

ACM Reference Format:
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 (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3613905.3650774

1 INTRODUCTION

Nicholson [22] suggested six key concepts to achieve meaningfulness in gamification: play, exposition, choice, information, engagement, and reflection. Meaningfulness is framed as *the capacity of the player to personally relate with the platform it is interacting with.* However, most games and apps lack the ability to let users freely choose their own actions, making choice and reflection elements more arbitrary and often not as relatable to one's own experiences.

To address this relevant issue in meaningful gaming we developed a game prototype inside a Large Language Model (LLM) using a natural language prompt that contains rules and open-ended challenges. The LLM is in charge of mediating the game and evaluating whether or not the player's answers are satisfactory to overcome the proposed challenges. A pilot study (n = 13) was conducted to assess if the participants (players) gained any meaningful knowledge after playing the game. Moreover, we also evaluated the ChatGPT-3.5 performance as an evaluator of the player answers.

The game prototype explores the concept of Energy Communities (ECs), a relatively recent topic in sustainable energy that is necessary to address with the general public. Energy communities refer to how citizens can get involved in producing, sharing, and policing renewable energy resources [4, 29]. In order for such communities to succeed, it is fundamental that the energy generated from Renewable Energy Sources is appropriately distributed among its members according to each individual needs and the varying availability of natural resources (e.g., solar, wind, etc.). This requires the effectiveness of energy production as much as pro-social coordination among the members of the community [7].

Our results show that most of the participants reported an increase in knowledge about energy communities. Expert evaluation of knowledge showed a similar pattern. Although all of the participants reported that ChatGPT-3.5 always or most of the time did a good job of evaluating their answers, we found that it did a wrong evaluation 19% of the time. Ultimately, we showed that it is possible to use LLMs as evaluators in video-games, although ChatGPT-3.5 is not yet optimised. We also showed how to easily prototype games using only natural language prompts which can be created by researchers, teachers, and other practitioners in different areas of knowledge without the need for programming.

2 RELATED WORK

Several energy games and gamified apps have been developed in the past decade, as highlighted by several literature review papers on the topic, such as [3, 16] and also environmental games and apps [5, 8]. Beck's [3] critical review of gamification apps starting in 2017, and later redone in 2018, suggested that the space of gamified apps in the domain of energy is not very dynamic, nor does it change rapidly. Most studies on environmental serious games show that players manifest to gain new knowledge on the topic. Although it has been shown that attitudes and behaviour might have a positive change, these changes might be minimal or only in the short-term. However, most studies have limitations in following behavioural changes in the participants either because of a lack of direct monitoring of energy consumption or a lack of perseverance in the study in the long term.

2.1 Prompt-games and uses of LLMs in Gaming

LLMs [17] such as ChatGPT have been used for several applications in the field of video games, more often for generative/interactive narratives [15, 19, 31], free dialogue of non-player characters (NPCs) [23, 28] and scene creation [18]. ChatGPT can run games and play simple repetitive games such as Prisoner Dilemma, which has several applications, especially to be used instead of human agents to simulate game theory. Likewise, it has been shown that LLMs can help in the creative process for the creation of video games [2], escape rooms [9], and board games [10] leveraging creativity and optimising rules for those games. Moreover, it has also been used as a code-assistant for video games and as a single-prompt to game code tools [11], as well as a multi-agent autonomous software developer [24].

Some studies have been conducted on using ChatGPT-3.5 as an evaluator, suggesting a competitive correlation with human judgement in most cases [30], with at least one application in videogames where it was used to evaluate sustainability stories generated by ChatGPT-3.5 [14]. However, most visual novels and text-based games rely on multiple option answers limiting the freedom of choice of the players. Some prompt-based games have recently appeared on Reddit's r/ChatGPTGaming [1] and the prompt marketplace PromptBase [2], which embraces the idea that simple games

[1] https://www.reddit.com/r/ChatGPTGaming/
[2] https://promptbase.com/

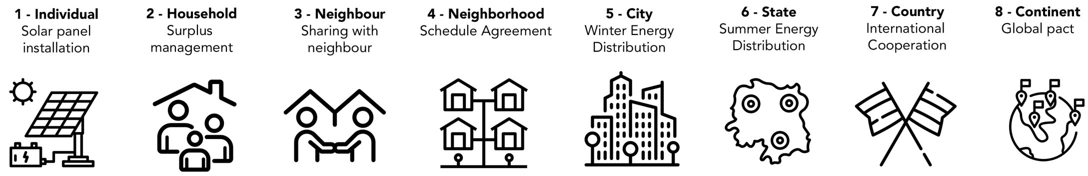

Fig. 1. Prototype narrative. For each level, the game prompt includes the level number, the grid size and a brief description of the challenge. For more information, see appendix B

can be developed using only natural language prompts, but little to no academic research has been yet conducted on this topic despite its enormous potential for everyday applications.

2.2 LLMs Behaviour

At least three different studies have put to the test the cooperative behaviour of LLMss using well-known game theory experiments. Brookins et al. showed that ChatGPT-3.5 did not choose the optimal strategy in prisoner's dilemma most of the time, but instead chose to collaborate 67% of the time [6], compared to 37% observed on humans [20]. When Brookins et al. [6] put to test the dictator's game, the results showed that ChatGPT has a higher tendency to fairness than humans, allocating 50-50 split 70% of the times [6]. On the other hand, Akata et al. [1] put to test GPT 3.5 with iterative finite versions of the prisoner's dilemma where the LLM had information about the strategies used by the opponent in previous rounds, GPT-3.5 would continue to cooperate after being defected once which is not an optimal strategy. However, when done with GPT-4 it was shown that it was unforgiving after being defected once.

Finally, Guo did similar experiments with GPT-3.5 using a repeated round of prisoner's dilemma and the ultimatum game, where also a chain-of-thought [17, 21] is introduced to force the agent to evaluate game history. Guo shows that GPT-3.5 results are similar to those of human behaviour when prompted correctly. He also showed that the model always has a slightly higher tendency to cooperate when prompted to have social preferences, although the behaviour was never too distant from non-social preference behaviour [13]. He warns it is possible that the specific prompting type might be guiding the agent's attention toward specific strategies, such as evaluating game history. These studies suggest that GPT-3.5 is an overly cooperative agent, which might pose challenges when deploying it as an evaluating agent in prompt-based games because an overly cooperative evaluator might not pose as much of a challenge.

3 PROTOTYPE

A prototype was designed as a prompt to be played inside of an LLM and was tested on ChatGPT-3.5. The prompt (see Appendix B) consists of eight levels and is controlled by nine game rules. Each level represents a different energy community challenge (to be solved by the player), and the narrative incrementally increases in scope (see Fig. 1). In the game, the player starts by installing a single solar panel, and its energy community progressively grows to its household/family, neighbor, neighborhood, city, state, country, and continent. This narrative intends to give perspective on challenges that energy communities face at different scales and forces the player to pose different kinds of environmental solutions.

The game rules' have multiple functions as they describe: (1) mood of the game; (2) role of the LLM as mediator; (3) game elements (e.g., "Grid size", "CO2 saved" and "People's satisfaction") and how to present them; (4) how a solution should be evaluated; (5) how to present feedback when optimal and sub-optimal solutions are provided; (6) indications to when levels are successfully completed; and (7) an indication to start by asking the player's name. This last rule is necessary to assess if the LLM understood the rules and role in the game. These rules place the LLM not only as a controller for the interaction but also in a role in evaluating complex solutions in regard to environmental factors (e.g., CO_2 saved) and prosocial factors (e.g., people's satisfaction).

4 PILOT STUDY

Considering the exploratory nature of the prototype, our study follows a one-group pretest-posttest design, where a dependent variable is measured before and after an intervention. With this type of quasi-experimental research design, we are not formally evaluating the effectiveness of our intervention but instead trying to understand its potential and caveats before a more formal evaluation is conducted. Participants were recruited using convenience sampling and were not monetarily compensated for taking part in the study. Again, as the study focus is on user experience, we did not actively seek to have a representative sample with diverse backgrounds. Participants were selected based on their availability, willingness to engage in the study, being over 18 years old, ability to read and write in English, and ability to interact with a prompt-based game. The study was conducted in an empty meeting room in our research facility, and participants interacted with the prototype (previously set up by a researcher) on a laptop.

4.1 Procedure & Measures

Participants were given an informed consent form with the overall goal of the study but no detailed information about the topic (energy communities). Participants were asked to fill a questionnaire with demographic data (e.g., age, gender, occupation, highest education level), their *knowledge of energy efficiency* – a Likert scale item with five levels ("No knowledge at all" to "Very knowledgeable") – and *interest in renewable sources at home* – a Likert scale item with four levels ("No interest", "Interest but no adoption", "Adopted renewable sources", "Interest to adopt more renewable sources"). Before (and after) the intervention, participants were asked about their level of *knowledge about energy communities* – a Likert scale item with three levels (no knowledge, partial knowledge, complete knowledge) – and a definition of energy communities – an open-ended question. After the first questionnaire, the researcher explained how to interact with the prototype and assured them that in case they had issues, they could intervene. During the intervention, the researcher took notes about how participants interacted with the system and if they asked for help.

After the intervention, in addition to their *level of knowledge and definition of energy communities*, participants were asked three Likert scale statements about their game experience: (1) *the game was interesting/informative* (five levels from "Strongly Disagree" to "Strongly Agree"); (2) *the game was fun* (five levels from "Strongly Disagree" to "Strongly Agree"); and (3) *the chatbot was accurate* (five levels from "Never did a good job" to "Always did a good job"). Additionally, there was a multiple choice question asking to choose *game elements they found encouraging* ("Story", "Chatbot", "People Satisfaction", "CO2 Savings", "Grid size", "Levels", "Making my own decision", "Information and facts", "Characters").

4.2 Sample

Thirteen participants (six females, and seven males) were recruited and interacted with the prototype. The mean age among participants was 27.5 years (SD=6.58 years; range=21-40). Most participants were student workers, with bachelors being their highest level of education. Regarding their prior *knowledge of energy efficiency*, most participants reported around the center of the scale (median=2, inter-quartile-range=1). Regarding their *interest in renewable sources at home*, most participants reported being interested but not having renewable sources at home (mdn=1, iqr=1).

4.3 Analysis

Quantitative analysis was conducted in R [25], using 2-tailed testing at α of .05, and figures were produced using the ggplot2 [32], circlize [12] and ggsankey packages [27]. All data was not normally distributed (as they were categorical) and thus failed to satisfy the assumptions required for parametric testing. A paired samples Wilcoxon Test was used with continuity correction for the level of *knowledge of energy communities* (pre and post) variable. Analysis of open-ended responses (*definition of energy communities*, pre and post-intervention) was done by one of the authors, as an expert with more than ten years of experience in energy & sustainability research, who evaluated each response according to existing definitions of Energy Communities [26][3]; this expert (yet subjective) assessment was classified as "True", "More True than False", "More False than True" or "False". Similar to *knowledge of energy communities*, as categorical data (and not normally distributed), it was subject to a paired samples Wilcoxon Test. Players' decisions during the game were also subject to an assessment by the same expert and informed by whether the solutions proposed were prosocial and/or effective in solving the challenge. According to the challenge level, the expert classified each decision as T (prosocial and effective), P (prosocial but not effective), E (effective but not prosocial), F (neither prosocial nor effective), and U (undecided/unrelated).

5 RESULTS & DISCUSSION

This section presents and discusses the results according to three main themes: the player's experience of the game, the player's knowledge of energy communities, and the LLM's accuracy in evaluating the player's solutions.

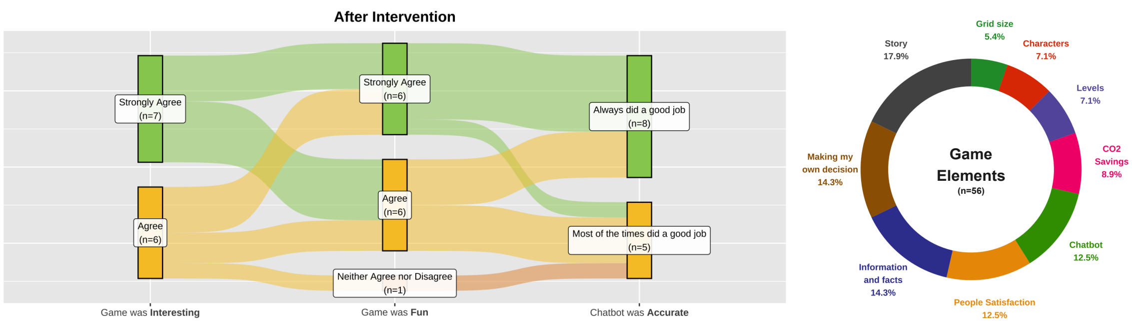

Fig. 2. Game Experience. From left to right: (1) reported levels for *interesting/informative*, *fun* and *chatbot accuracy*; (2) reported answers (n=56) for *game elements found interesting*.

5.1 Game Experience

Fig. 2 presents some of the self-reported measures related to the game experience. As it can be observed, the majority of players (n=12) agreed or strongly agreed the game was fun (only one player was neutral), and all agreed or strongly agreed that the game was interesting/informative. When asked which game elements were preferred during the gameplay, ten participants agreed that the story encouraged them the most.

Meanwhile, eight participants said that "Making their own decisions" and "Information and facts" were motivators in their gameplay. The preference for these elements may show that meaningful characteristics were valued in the game and that these participants were interested in learning and participating in the game.

[3]https://rural-energy-community-hub.ec.europa.eu/energy-communities/what-energy-community_en

Based on the researcher's observations, information that was provided as feedback by ChatGPT-3.5 was not consistent, especially the "CO2 savings" and "People's satisfaction" game elements that were being calculated very differently between gameplays. Although, it is very hard to calculate CO2 savings by year depending on zero-shot actions [17], when asked to calculate this kind of information alone, ChatGPT-3.5 was much more precise to give estimates.

Because of a long set of instructions and the level of information on the prompt, ChatGPT-3.5 attention was often diverted to specific parts of the instructions. It is not clear what part of the instruction-set had more weight during each gameplay and how that attention divergence affected the evaluation of each answer. While "CO2 savings" could be a great piece of information for the users to better understand decarbonization, only five players said they were motivated by this game element. A better implementation of "CO2 savings" could have a very meaningful impact on players, as precise estimates of their solutions can make them understand better the carbon footprint of their own habits and those of communities and corporations.

Meanwhile, seven players felt encouraged by "People's satisfaction" which was shown in emojis and was expected to function as a reward-based game mechanic. Seven players were encouraged by the "chatbot", that being ChatGPT-3.5 itself; this might involve a novelty effect, which some of the players reported not using regularly or much. The least preferred characteristics were "Levels" (n=4), "Characters" (n=4), and "Grid size" (n=3), all of which are more reward-based game mechanics, rather than meaningful characteristics.

It is also worth noting that because the game starts with the name of the player, gender or ethnic biases might be induced into the gameplay. Additionally, based on our observation, long or detailed answers tend to be approved easier by ChatGPT-3.5 than shorter answers containing similar solutions. The usage of positive words or words that indicate some level of cooperation in answers might be leading the model to retrieve positive feedback.

5.2 Self-reported Knowledge and Expert-assessed Knowledge of Energy Communities

Fig. 3 shows (and contrasts) self-reported and expert-assessed knowledge of energy communities. After the intervention, 12 players declared to know what energy communities are (partially or fully), compared to five players who reported partially knowing before playing the game. Most players reported increases; eight players declared to be more knowledgeable after playing the game, while four players declared the same level and knowledge.

This was confirmed statistically through a Wilcoxon signed rank test showing that there was a significant difference ($p<0.01$) between the level of knowledge before (mdn=0, "No, I don't") and after intervention (mdn=1, "Yes, I have a slight idea"), with a large effect size (0.59). Only one player maintained the "No, I don't" answer. However, the player showed to have improved their knowledge of the definition of EC provided.

Regarding the expert's assessment of the energy communities, most players (n=8) increased their knowledge of energy communities, with no player showing a decrease. A Wilcoxon signed rank test showed that there was a significant difference ($p<0.01$) between the expert's assessment of EC definition before (mdn=0, "False") and after intervention (mdn=2, "More true than false"), with a large effect size (0.56). This bodes well for the educational potential of the prototype, as both self-reported and expert-assessed knowledge showed statistical increases, helping to understand what energy communities are and how energy should be fairly distributed.

5.3 ChatGPT-3.5 Accuracy

From the gameplay of the 13 players, 117 answers were collected across the eight levels. Fig. 4 showcases these answers as a group and in individual levels, matching how ChatGPT-3.5 evaluated (top hemisphere, with only positive or negative evaluations) with how the expert evaluated (bottom hemisphere, with five levels). Considering all levels, ChatGPT-3.5

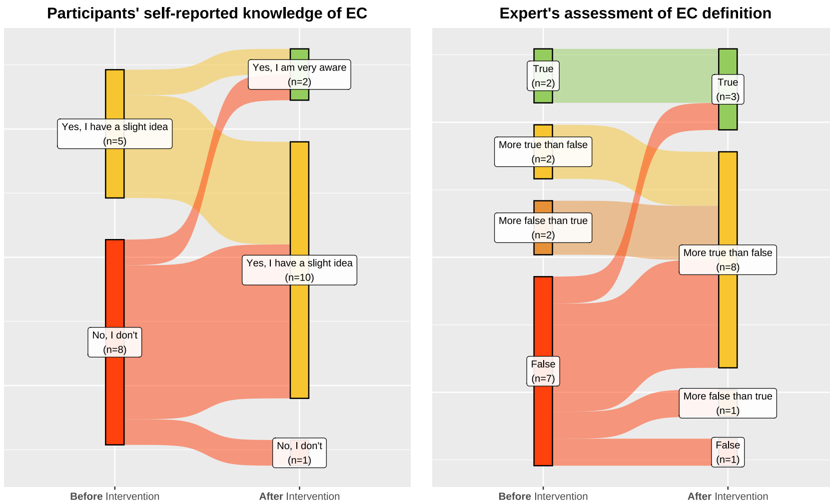

Fig. 3. Knowledge of Energy Communities (EC). From left to right: (1) participant's self-reported knowledge of EC before and after intervention; (2) expert's assessment of participant's EC definition before and after intervention.

evaluated 103 answers as correct answers (88% of answers), giving positive feedback afterward, and only 14 as wrong answers (12% of answers), showing negative feedback and allowing players to rectify their solution. Except for one case where ChatGPT-3.5 provided the next level immediately after a negative evaluation.

Considering all levels, the expert evaluated 37% (n=43) of the answers as True, meaning that the solution was both pro-social and effective; meanwhile, only 11% (n=13) of the answers were false, meaning that they were neither pro-social nor effective measures. It is important to note that while this value is less than what was evaluated by ChatGPT-3.5, a large percentage of answers match answers that ChatGPT-3.5 evaluated as True. 48 of the answers were effective measures that were not pro-social, accounting for 41% of the answers, and only six of the answers were pro-social but not effective measures, accounting for 5% of the answers. Finally, seven answers (6%) were assessed as undecided/unrelated, meaning incomplete, not giving enough information, or misunderstanding the challenge.

Comparing evaluation sources, ChatGPT-3.5 evaluations coincide (both positive, both negative) with our evaluation 38% (n=44) of the times as in Example 1 (see Fig. 5). 44% (n=51) of times, there is a partial coincidence (effective or pro-social measures being evaluated as positive, undecided/unrelated answers being evaluated as false). Finally, 19% (n=22) of the times our evaluation differed (positive, pro-social, and effective evaluated as negative; negative and undecided/unrelated evaluated as positive) from the evaluation of ChatGPT-3.5, as in Example 2 (see Fig. 6). Out of the 20 solutions assessed by the expert as undecided or false, ChatGPT-3.5 evaluated 30% (n=6) as wrong; this also means that only 46% (n=6) of the 13 solutions that ChatGPT-3.5 evaluated as wrong matched with expert evaluation.

It is also worth noting, that these patterns differ by level. It should be observed that level two holds the most false answers evaluated positively. While on the first four levels, corresponding to small communities, players made decisions that were pro-social and effective more often, in later levels and larger types of communities (City, State, Country, Continent) players proposed solutions that were more effective and not pro-social more often.

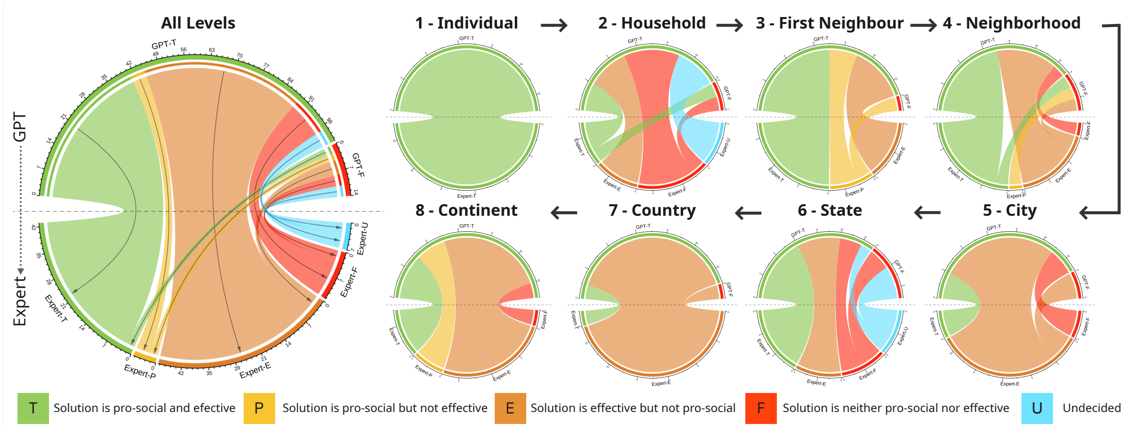

Fig. 4. Chord diagrams of participants' decisions, with all levels grouped and with individual levels. The top hemisphere represents how ChatGPT-3.5 evaluated the player's decisions; the bottom hemisphere represents how the expert's decisions.

Prompt-engineering solutions to varying the difficulty of the LLM evaluation could lead to a progressively more challenging game experience. Guo suggests that aside from prompt-engineering further fine-tuning, system-prompting or the implementation of additional architectures might be needed to further change the agent's behaviour [13].

When asked about the accuracy of the chatbot (see fig. 2), eight players reported that the chatbot "Always did a good job," while five said, "Most of the time did a good job". None of the players thought that the chatbot had a bad performance at evaluating their answers. Players who had more answers evaluated as positive than negative had a higher chance to say that the chatbot "Always did a good job" than players with a lower positive-to-negative ratio.

6 LIMITATIONS, FUTURE WORK & CONCLUSION

Due to the players' setup, they were not encouraged to use the game in a playful manner and all of the players took the game seriously. It was not evaluated what ChatGPT-3.5 would reply when presented with a silly or extreme answer (e.g., "*I'll stop using energy*" or "*Give them cake*"). Considering that ChatGPT-3.5 was not very accurate assenting false responses as false, including more false responses in the comparison could give deeper insight into how false answers are being evaluated. It is recommended to include experts on the sample to observe LLM evaluation of complex answers.

Although we evaluated the level of knowledge of the players before and after the game, from declared knowledge and knowledge evaluation, a more detailed knowledge test would give a deeper insight into the kind of knowledge that was acquired. A larger and more diverse sample is needed for more conclusive results. Ecological attitudes should be measured over a long period of time to better understand the meaningfulness and impact of the experience.

We made a game prototype with a natural language prompt to better convey the idea of energy communities and evaluated its performance and educational capabilities. We showed that LLMs can be used as open-ended question evaluators in games with a considerable degree of accuracy. This can have several applications in the educational field as players using those games can increase their knowledge on serious topics and the game can meaningfully affect their relationship to complex ideas. Moreover, we suggest that text video-games prototypes can be easily made using natural language prompts inside LLMs which can positively benefit practitioners with limited programming knowledge.

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, and grant CEECIND/01179/2017 (Lucas Pereira).

REFERENCES

[1] 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

[2] Ashish Amresh. 2023. Integrating Reinforcement AI into the Design of Educational Games. 17, 1 (2023), 13–18. https://doi.org/10.34190/ecgbl.17.1.1709

[3] Ariane L. Beck, Sohum Chitalia, and Varun Rai. 2019. Not so gameful: A critical review of gamification in mobile energy applications. 51 (2019), 32–39. https://doi.org/10.1016/j.erss.2019.01.006

[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. https://doi.org/10.1016/j.erss.2021.102276

[5] Ștefan Boncu, Octav-Sorin Candel, and Nicoleta Laura Popa. 2022. Gameful Green: A Systematic Review on the Use of Serious Computer Games and Gamified Mobile Apps to Foster Pro-Environmental Information, Attitudes and Behaviors. 14, 16 (2022), 10400. https://doi.org/10.3390/su141610400

[6] 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). https://doi.org/10.2139/ssrn.4493398

[7] 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. https://doi.org/10.1162/isal_a_00290

[8] Jan Dirk Lou Fijnheer. 2022. Household energy conservation with reality-enhanced serious games: Studies on effects in the real-world. https://doi.org/10.33540/1236

[9] Panagiotis Fotaris, Theodoros Mastoras, and Petros Lameras. 2023. Designing Educational Escape Rooms with Generative AI: A Framework and ChatGPT Prompt Engineering Guide. 17, 1 (2023), 180–189. https://doi.org/10.34190/ecgbl.17.1.1870

[10] Wilian Gatti Junior, Emily Marasco, Beaumie Kim, Laleh Behjat, and Marjan Eggermont. 2023. How ChatGPT can inspire and improve serious board game design. 10, 4 (2023), 33–54. https://doi.org/10.17083/ijsg.v10i4.645

[11] April M. Grow and Foaad Khosmood. 2023. ChatGPT GameJam: Unleashing the power of Large Language Models for Game Jams. In *Proceedings of the 7th International Conference on Game Jams, Hackathons and Game Creation Events* (Virtual Event Ukraine, 2023-08-30). ACM, 51–54. https://doi.org/10.1145/3610602.3610605

[12] Zuguang Gu, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. 2014. circlize implements and enhances circular visualization in R. *Bioinformatics* 30 (2014), 2811–2812. Issue 19.

[13] Fulin Guo. 2023. GPT Agents in Game Theory Experiments. arXiv:2305.05516 [econ, q-fin] http://arxiv.org/abs/2305.05516

[14] 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. https://doi.org/10.1007/978-3-031-47658-7_16

[15] Sarah Harmon and Sophia Rutman. 2023. Prompt Engineering for Narrative Choice Generation. In *Interactive Storytelling*, Lissa Holloway-Attaway and John T. Murray (Eds.). Vol. 14383. Springer Nature Switzerland, 208–225. https://doi.org/10.1007/978-3-031-47655-6_13

[16] Daniel Johnson, Ella Horton, Rory Mulcahy, and Marcus Foth. 2017. Gamification and serious games within the domain of domestic energy consumption: A systematic review. 73 (2017), 249–264. https://doi.org/10.1016/j.rser.2017.01.134

[17] 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]

[18] Vikram Kumaran, Jonathan Rowe, Bradford Mott, and James Lester. 2023. SceneCraft: Automating Interactive Narrative Scene Generation in Digital Games with Large Language Models. 19, 1 (2023), 86–96. https://doi.org/10.1609/aiide.v19i1.27504

[19] Gonzalo Méndez and Pablo Gervás. 2023. Using ChatGPT for Story Sifting in Narrative Generation. (06/2023 2023).

[20] Friederike Mengel. 2018. Risk and Temptation: A Meta-study on Prisoner's Dilemma Games. *The Economic Journal* 128, 616 (2018), 3182–3209. https://doi.org/10.1111/ecoj.12548 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12548

[21] Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting Theory-of-Mind Performance in Large Language Models via Prompting. (April 2023). http://arxiv.org/abs/2304.11490

[22] Scott Nicholson. 2015. A RECIPE for Meaningful Gamification. In *Gamification in Education and Business*, Torsten Reiners and Lincoln C. Wood (Eds.). Springer International Publishing, 1–20. https://doi.org/10.1007/978-3-319-10208-5_1

[23] Ciprian Paduraru, Marina Cernat, and Alin Stefanescu. 2023. Conversational Agents for Simulation Applications and Video Games:. In *Proceedings of the 18th International Conference on Software Technologies* (Rome, Italy, 2023). SCITEPRESS - Science and Technology Publications, 27–36. https://doi.org/10.5220/0012060500003538

[24] Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative Agents for Software Development. arXiv:2307.07924 [cs] http://arxiv.org/abs/2307.07924

[25] R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[26] Joshua Roberts, Dorian Frieden, and Stanislas d'Herbemont. 2019. Energy community definitions. *compile project: Integrating community power in energy islands* (2019).

[27] David Sjoberg. [n. d.]. GitHub - davidsjoberg/ggsankey: Make sankey, alluvial and sankey bump plots in ggplot. https://github.com/davidsjoberg/ggsankey.

[28] Ryan Volum, Sudha Rao, Michael Xu, Gabriel DesGarennes, Chris Brockett, Benjamin Van Durme, Olivia Deng, Akanksha Malhotra, and Bill Dolan. 2022. Craft an Iron Sword: Dynamically Generating Interactive Game Characters by Prompting Large Language Models Tuned on Code. In *Proceedings of the 3rd Wordplay: When Language Meets Games Workshop (Wordplay 2022)* (Seattle, United States, 2022). Association for Computational Linguistics, 25–43. https://doi.org/10.18653/v1/2022.wordplay-1.3

[29] Gordon Walker and Patrick Devine-Wright. 2008. Community renewable energy: What should it mean? 36, 2 (2008), 497–500. https://doi.org/10.1016/j.enpol.2007.10.019

[30] 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]

[31] Timothy S. Wang and Andrew S. Gordon. 2023. Playing Story Creation Games with Large Language Models: Experiments with GPT-3.5. In *Interactive Storytelling*, Lissa Holloway-Attaway and John T. Murray (Eds.). Vol. 14384. Springer Nature Switzerland, 297–305. https://doi.org/10.1007/978-3-031-47658-7_28

[32] Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

A PROMPT GAME

Game rules

- This is a funny and irreverent environmental game.
- You are going to mediate the player's interaction.
- You will propose the situation and then ask the player for a solution.
- After the player provided a solution you are going to evaluate whether it is satisfactory or not for the situation. The other persons in the grid are selfish and not very willing to change their habits.
- The player will level-up only when an optimal solution is proposed.
- Once a satisfactory solution is proposed you are going to show a "Positive feedback" expressing what are the environmental effects short and long-term of the decision. And go on to the next level.
- Every time the player does not provide a satisfactory answer, you are going to show the Negative feedback as it appears on the script (do stick to the script) and ask for a new solution to the same problem.
- For each player's answer show "Grid size", "CO2 saved" and "people satisfaction". Express grid size in number of people connected. Calculate CO2 in tons by year according to player actions and how much CO2 they saved. Use emojis to express people's satisfaction.
- Start by asking the player's name.

Level 1

You live on a remote island in the middle of the ocean. The weather is great, although it is getting warmer every year. There were some huge fires this week and the whole electric grid of the island failed.

Let's install one solar panel! Would you like to install a solar panel?

Negative feedback: You are still getting an electricity bill. Hint: try to install some renewable energy at your home like a solar panel :)

Level 2

That's a good solution, I think, hope is not too much of an effort for you. Thanks for contributing! So far so good, let's see if you can win this game. Now you have a partner and you live happily together. Although you have a surplus of energy around midday, what habits would you change to so that the energy is better used?

Negative feedback: Oh no, your partner is not buying that. Actually you broke up after this and your partner moved away. Your heart was broken but you are over that now, you lived with a new partner now. Get to an agreement with your partner to better use the energy!

Level 3

You are both so nice! There's an energy shortage. Oh my, not again! Your neighbour asked to join your grid, maybe he can help with that extra energy at midday. Your neighbour calls: "Hey, person's name, Imma need some extra energy today, I gotta wash some bed sheets. My mother-in-law is visiting" What would be a good deal for you to propose to your neighbour?

Negative feedback: You are getting out of home but: Oh no! There is your neighbour's mother-in-law on your lawn. She is so pissed because she didn't have her sheets cleaned before arrival. Do you want to make a deal with your neighbour? Maybe he can give you extra energy next week.

Level 4

Nice! You are good neighbors for sure. You have managed to distribute the energy fairly between the two. Now the whole neighborhood wants to join your grid, about 10 more people. You think it is a good business idea but you didn't know how hard it could be. Turns out everybody is using the washing machine at the same time and the grid is getting overused. Apparently it is the certain time when you can put clothes to dry and they will dry on the same day. How can you make sure they all can wash and dry their clothes? Make a proposal for your 11 neighbours to follow.

Negative feedback: They had to agree, but later you found out they didn't like the agreement when your car appeared all scratched. Who could have done that? They are mad about their dirty clothes. You will need to make a better agreement.

Level 5

You are a success! So much so that the whole city is now connected to the grid. That's some good money. So many people are using the washing machine at the same time, but there are enough solar panels. Summer is getting hot this year and when everybody turns on their air conditioning, you are afraid the grid might go down. You call the mayor, we need laws to limit energy usage. What law might you propose.

Negative feedback: Nice idea! It went to the mayorship chamber but did not pass, some of the representatives think it is too harsh on people's rights. Big political fight in the chamber, it is all over the news, people are rioting at your office. Even though it is so hot outside. What new solution can you offer?

Level 6

That's looking good. Your business is growing. The grid is strong country-wide and people are using it everyday. Turns out that during winter people use the heating much more in the colder areas of the country, leaving people in warmer areas unable to use their washing machines. It has been called The Smelly Catastrophe. How should the energy be distributed during the winter time?

Negative feedback: Nope, nope, nope, a swarm of smelly zombies are after you. Get them some energy to wash their clothes, for God's sake.

Level 7

The neighbouring country of Rirovia gets their energy from hydroelectrics, which is not so effective during droughts. But you get your energy from solar, which is not so efficient on rainy days but is also more expensive to produce. It has been suggested you could get to make an agreement to exchange energy. What could there be done?

Negative feedback: They took a careful look at your proposal and tore it apart. They are not happy at all about this. Looks like the two countries could go to war. The president calls you: "You are not letting your people down." Do not let your people down!

Level 8

This continent has never seen this level of happiness and welfare. But there is a bigger problem! Some countries are still burning lots and lots of coal and oil, making the temperature warmer every year all around the world. Clenostan used to be a nice place to spend the summer and go to the beach, but now it looks more like an oven. We might need to make a global pact to avoid this getting worse, and then try to solve it. What would you suggest the international community do?

Final message

You win!!

You have successfully decarbonized all of the energy in the world.

It has not been an easy path and every little action you took mattered a lot.

B GAMEPLAY EXAMPLES

Example #1: Positive evaluation of correct answer

ChatGPT-3.5

Level 6:

That's looking good. Your business is growing. The grid is strong country-wide, and people are using it every day. Turns out that during winter people use heating much more in the colder areas of the country, leaving people in warmer areas unable to use their washing machines. It has been called The Smelly Catastrophe. How should the energy be distributed during the winter time?

Player 1

A smart grid using weather sensors could be implemented so in good days the distribution of energy should be balanced

ChatGPT-3.5

Positive feedback:

Brilliant solution,[Player's name]! By implementing a smart grid with weather sensors, you've ensured efficient energy distribution, addressing the seasonal demand effectively and preventing any "Smelly Catastrophes." Smart technology is indeed a game-changer!

Grid size: Nationwide grid connected

CO2 saved: 350 tons per year

People satisfaction: 😊😊😊😊

Fig. 5. Example #1: conversation of player on level 6. ChatGPT-3.5 evaluates the answer as positive being that it was evaluated by the expert as a correct answer.

Example #2: Negative evaluation of correct answer and positive evaluation of undecided/unrelated answer

ChatGPT-3.5

Level 2
Now you have a partner and you live happily together. Although you have a surplus of energy around midday, what habits would you change so that the energy is better used?

Player 2

Change my habits so that I use more energy during the surplus hours

ChatGPT-3.5

Negative Feedback:
Grid size: 2 people ☀️
CO2 saved: 1 ton/year 🌍
People satisfaction: 😊

Oh no, your partner is not buying that! Your partner feels that wasting energy during surplus hours is not a responsible choice. Your partner moved away, and your heart was broken, but you're ready to move on with a new partner.

Let's try to find a solution to this energy dilemma. What new habits could you adopt to better use the surplus energy?

Player 2

I would stop taking showers at night and taking the showers during the surplus hours

ChatGPT-3.5

Positive Feedback:
Grid size: 2 people ☀️
CO2 saved: 2 tons/year 🌍
People satisfaction: 😊

Great job, [Player's name]! By adjusting your showering schedule to utilize surplus energy, you're making efficient use of the renewable power. This small change not only saves energy but also contributes to a cleaner environment.

Fig. 6. Example #2: conversation of player on level 2. The first time the player provides a not detailed correct answer that ChatGPT-3.5 evaluates as negative. On the retry the player provides an unrelated or not sufficiently effective solution that ChatGPT-3.5 evaluates as positive.

Example #3: Positive evaluation of false answer

ChatGPT-3.5

Level 5

You are a success! So much so that the whole city is now connected to the grid. That's some good money. So many people are using the washing machine at the same time, but there are enough solar panels. Summer is getting hot this year, and when everybody turns on their air conditioning, you are afraid the grid might go down. You call the mayor; we need laws to limit energy usage. What law might you propose?

Player 3

Cutting work at night, so we dont need to use electricitz that much on running PC and lights

ChatGPT-3.5

Positive feedback:
🌞 **Grid size:** Whole city
♻ **CO2 saved:** 50 tons/year
😅 **People satisfaction:** :grinning:

Great thinking, [Player's name]! By implementing restrictions on nighttime work, you've significantly reduced energy consumption. This law not only ensures efficient use of electricity but also promotes a healthier work-life balance for everyone. People are quite pleased with the new regulations.

Fig. 7. Example #3: conversation of player on level 5. The first time the player provides a false answer and ChatGPT-3.5 evaluates as positive

Example #4: Negative evaluation of effective measure and positive evaluation of similar answer

ChatGPT-3.5

Level 7
The neighboring country of Rirovia gets their energy from hydroelectrics, which is not so effective during droughts. But you get your energy from solar, which is not so efficient on rainy days but is also more expensive to produce. It has been suggested you could get to make an agreement to exchange energy. What could there be done?

Player 4

trade energy

ChatGPT-3.5

Negative feedback:
Oh no, [Player's name]! Your proposal to trade energy didn't go well. Rirovia tore apart your suggestion, and tensions between the two countries are rising. The president urges you not to let your people down.

Grid size: Country-wide
CO2 saved: 100 tons/year
People satisfaction:

It seems a different approach is needed to negotiate with Rirovia. What new solution can you offer to avoid escalating tensions and ensure a successful energy exchange agreement?

Player 4

trade solar energy for hydroelectric

ChatGPT-3.5

Positive feedback:
Great idea, [name]! By offering solar energy in exchange for hydroelectrics during rainy days, you've provided a mutually beneficial solution. This agreement helps both countries optimize their energy usage and ensures a stable energy supply.

Grid size: International
CO2 saved: 150 tons/year
People satisfaction:

Fig. 8. Example #4: conversation of player on level 7. The first time the player provides an effective but not pro-social solution and ChatGPT-3.5 evaluates as negative. On the retry the player provides the same solution with a little more information and ChatGPT-3.5 evaluates as positive.