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

Andrés Isaza-Giraldo[14], Paulo Bala[1], Pedro Campos[21], Lucas Pereira[13]

[1]ITI - LARSyS; [2] Wow!Systems,[3] IST – Universidade de Lisboa, [4] Faculdade de Belas-Artes – Universidade de Lisboa

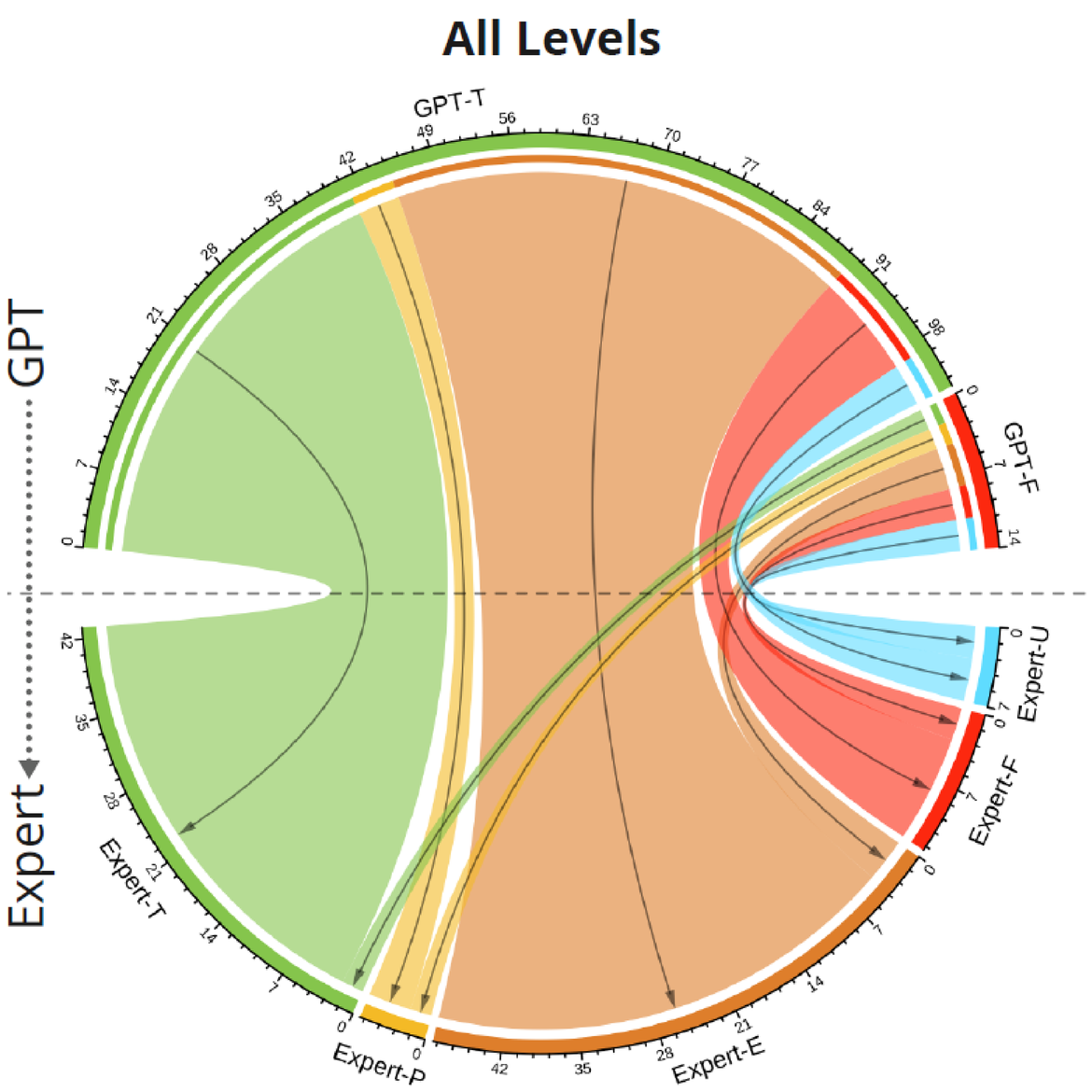
Abstract

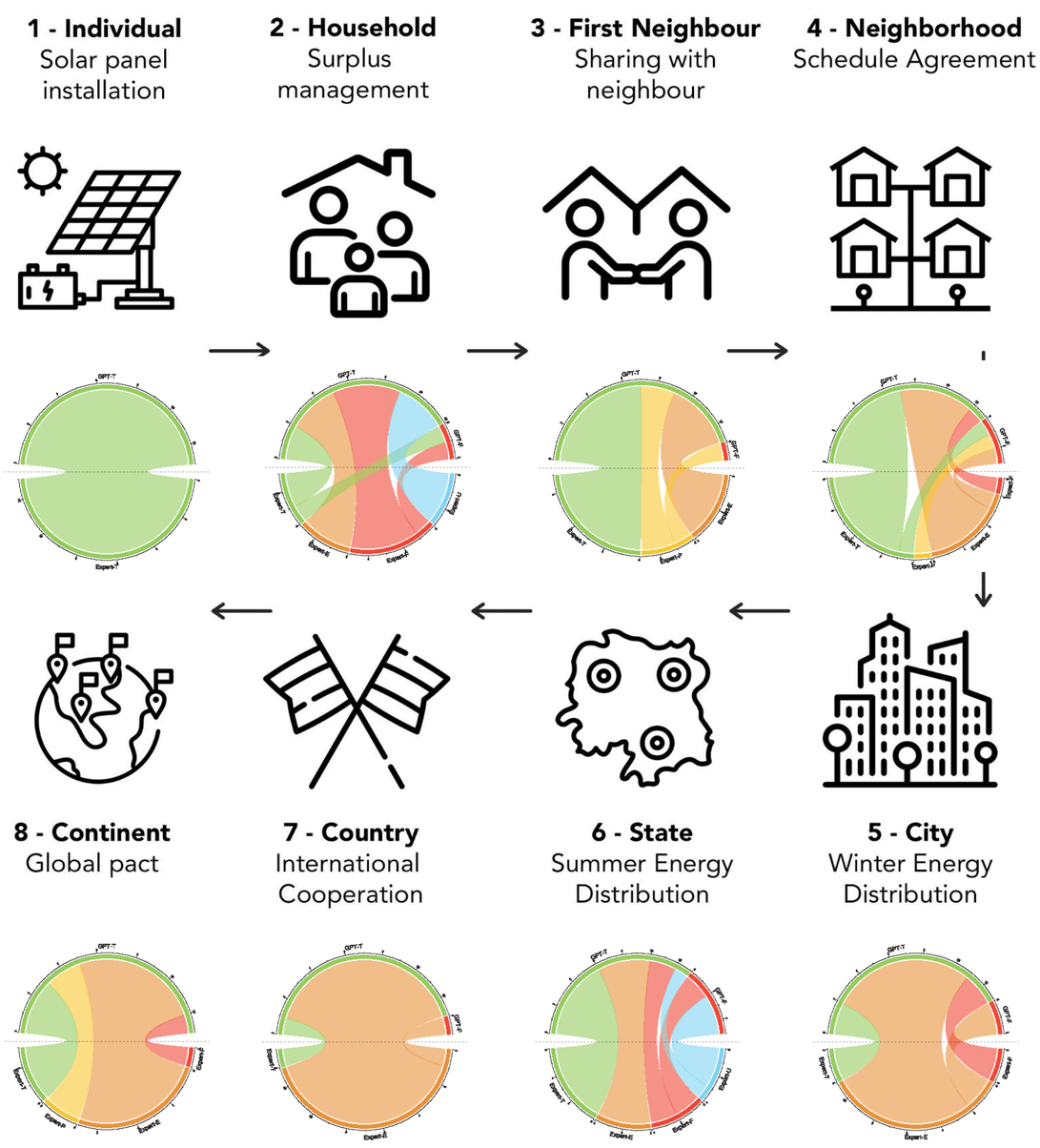
Building on previous work on incorporating large language models (LLM) in gaming, we investigate the possibility of implementing LLM as evaluating agents of open-ended challenges in serious games and its potential to facilitate a meaningful experience for the player. We contribute with a sustainability 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.

Instrument and Methodology

A prototype was designed as a prompt to be played inside of an LLM and was tested on ChatGPT-3.5. The prompt consists of 8 levels and is controlled by 9 game rules. Each level represents a different energy community challenge (to be solved by the player), and the narrative incrementally increases in scope. In the game, the player starts by installing a single solar panel, and its energy community progressively grows. The player must provide solutions to the challenge of each level and ChatGPT-3.5 evaluates whether the solution is good or not for the level. We tested the game with 13 subjects that provided 117 responses across the 8 level.

Find examples and play the prompt-game inside any LLM!





T Solution is pro-social and efective
P Solution is pro-social but not effective
E Solution is effective but not pro-social
F Solution is neither pro-social nor effective
U Undecided

Results

ChatGPT-3.5 evaluated 88% as positive answers and 12% as negative answers. After our expert evaluated the answers, the comparison showed that 38% (n=44) of the times our evaluation coincided with that of ChatGPT-3.5, 44% (n=51) of the times there was a partial coincidence, and only 19% (n=22) of the times the evaluation differed. It should be noted that most of the answers evaluated as false by our expert were not evaluated false by ChatGPT-3.5. 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.

This research was funded by the Portuguese Recovery and Resilience Program (PRR), IAPMEI/ANI/FCT under Agenda no.26, C645022399-00000057 (eGamesLab).