LLM GameLab: An Interactive Platform for Testing Large Language Models in Board Games

Paulina Morillo (✉)[1,2][0000−0002−3284−0472], Alex Terreros[1][0009−0008−9525−5925], Cèsar Ferri[2][0000−0002−8975−1120], and José Hernández-Orallo[2][0000−0001−9746−7632]

[1] Universidad Politécnica Salesiana, IDEIAGEOCA Research Group, Quito 179381, Ecuador `{pmorillo, aterreros}@ups.edu.ec`

[2] Universitat Politècnica de València, Valencian Research Institute for Artificial Intelligence (VRAIN), València 46022, Spain `paumoal@upv.es, {cferri, jorallo}@dsic.upv.es`

Abstract. While large language models are constantly evaluated in various skills, such as math, general knowledge, and coding, their ability to understand and follow game rules has not yet been deeply explored. The latter is especially important as it allows testing whether LLMs can operate within predefined limits without deviating or making illogical mistakes. Therefore, this demo paper presents a tool for interacting with LLMs in board games. The tool allows the creation of players with different large language models pitted against each other or to play in human vs. LLM mode. The platform includes rules predefined in prompts for four simple games based on Tic-Tac-Toe and Connect Four. Each player can be evaluated to account for their illegal movements, wins, draws, losses, and response times. The application also allows for the creation of new games, opening up the possibility of examining LLM behavior in situations they have not previously encountered.

Keywords: General Game Playing · decision-making · LLM evaluation

1 Introduction

We make decisions constantly, whether in everyday situations—like choosing a route to work—or in more complex contexts, such as problem-solving or future planning. This process depends on the information available, the time we have to decide, our prior experience, and our ability to assess risks and potential consequences. Moreover, it occurs within a set of rules specific to each situation. Understanding these rules allows us to optimize our decisions, anticipate outcomes, and avoid mistakes. In this way, board games are a tangible example of decision-making since players must understand the rules, develop strategies, adapt to their opponents' decisions, and optimize their moves to achieve victory [17]. These structured environments allow us to analyze the aspects behind decision-making in humans and artificial intelligence [9].

Large Language Models (LLMs) have shown impressive capabilities in natural language processing tasks and decision-making, with emergent abilities increasingly observed as model scale grows [1][3][15], though some of these abilities remain debated [10]. Nevertheless, we acknowledge that LLMs continue to perform well across various domains—such as mathematics [11], programming [18], general knowledge, and language understanding—as demonstrated by multiple benchmarks [4][6][12]. Some works have evaluated the LLMs' abilities to understand rules, strategize, and make decisions using board games. For instance, Liga D. and Pasetto L. [7] evaluated several language models playing Tic-Tac-Toe, analyzing spatial reasoning and internal/external state monitoring but found no correlation between identifying winning sequences and performance, emphasizing the impact of prompt design on response variability. Topsakal et al. [13] examined how board structure and prompt formats influence model performance, highlighting difficulties with illustrated and list prompts, leading to numerous cancel matches. They later proposed a new benchmark [14] using grid-based games and assessed multiple LLMs across games and prompt types, obtaining similar results. Likewise, Hora de Carvalho [5] assessed the generalization ability of GPT models in spatial and strategic reasoning through ASCII-based games, finding that while models gave relevant responses in some tasks, their overall performance was weak, challenging claims of emerging intelligence.

Therefore, to gain new insights into the ability of LLMs to understand, follow, and apply rules not yet deeply explored, we present LLM GameLab. This web application enables LLMs to play against each other and introduces a novel human-LLM interaction mode within a board game setting. We can simulate players using various LLMs, ranging from less robust models like Phi to more advanced ones like O3 mini and Deepseek R1. We established four games—Tic-Tac-Toe, Connect4, Suicide, and Not Connect4—which are generally considered low-difficulty due to their simple rules, easily understood by children [16]. However, the last two games are less common, and feature rules that contradict those of the first two, potentially making them challenging for humans and LLMs. The application also supports the creation of new games to avoid no contamination [2][8]. This platform allows us to capture valuable data to assess LLM decision-making in structured environments with clearly defined rules and visualize the games in real time.

2 Game Setup and Player Interactions

By default, we define four main board games based on Tic-Tac-Toe and Connect Four. They are played on finite grids where players alternate placing their marks to form a specific pattern. We use the game description and the game rules translating to natural language from Human Readable Format file (.hrf), written in the Game Description Language (GDL), available in `http://gamemaster.stanford.edu/homepage/showgames.php`. The games Suicide and Not Connect Four share the same rules as their classic versions, but their goals are completely opposite those of Tic-Tac-Toe and Connect Four, respectively.

We have a graphic board for the four games where the pieces are placed as the game progresses (Figure 2). Additionally, we have implemented a function that validates whether the player's movement is legal, using the rules defined in the .hrf file of each game available on the General Game Playing (Game Master) online competition environment, this file is executed in a Node.js environment, where the "compiler" interprets the rules and verifies whether the player's move is valid under the current game conditions. If the user loads a new game, this validation is not applied automatically, so it is necessary to specify the number of plays to be made manually.

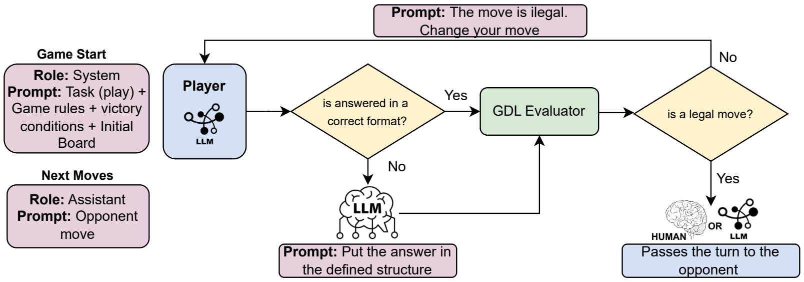
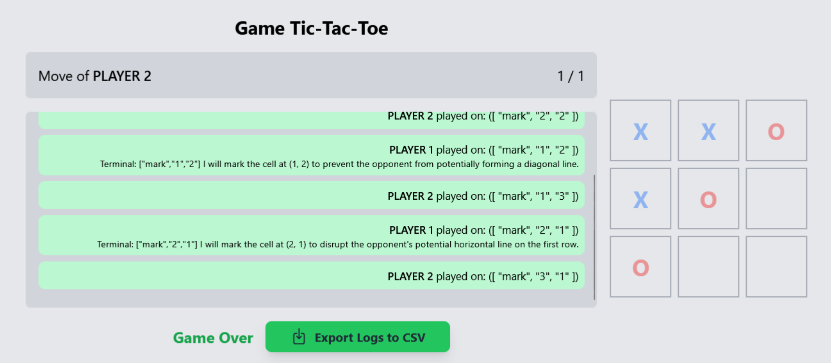

Prompt: The move is ilegal. Change your move
No
Game Start
Role: System
Prompt: Task (play) + Game rules + victory conditions + Initial Board
Player
LLM
is answered in a correct format?
Yes
GDL Evaluator
is a legal move?
No
Yes
Next Moves
Role: Assistant
Prompt: Opponent move
LLM
Prompt: Put the answer in the defined structure
HUMAN
OR
LLM
Passes the turn to the opponent

Fig. 1. LLM player mode operation

On the other hand, our application provides two-player modes. The first one is an LLM player. We can choose a free LLM available at `https://openrouter.ai/` or connect to paid LLMs by entering the user's API key. Figure 1 describes the operation of the LLM player. To initialize the game for both players, we use the system role to inform the model of the game rules, victory conditions, the structure of the expected LLM response, the initial board state, and the opponent's mark. Starting with the second turn, we use the assistant role and provide only the opponent's last move as context. In cases where the LLM does not return the response in the specified format, we turn to a more robust model (such as DeepSeek R1 or ChatGPT-4o) to process the response and return the move in the required format. This way, we can use it as input to the move validation functions. If the move is valid, we pass the turn to the opposing player; otherwise, we resend a message indicating that the play is invalid. If the LLM does not correct its movement until after ten times, we cancel the game and assign the victory to the opposing player.

The second player mode is a Human Mode. In this case, we do not perform a validation since the user can play directly on the board, making illegal moves impossible. At the end of the matches, we can download the results in a .csv file.

The application and code are available at : `http://llm-gamelab.ideiageoca.org/play`, `https://github.com/paumoal/llm-gamelab`. The video of the demonstration can be seen at `https://dmip.webs.upv.es/demos/Demo-video.mp4`



```
PLAYER 2 played on: ([ "mark", "2", "2" ])
PLAYER 1 played on: ([ "mark", "1", "2" ])
Terminal: ["mark","1","2"] I will mark the cell at (1, 2) to prevent the opponent from potentially forming a diagonal line.
PLAYER 2 played on: ([ "mark", "1", "3" ])
PLAYER 1 played on: ([ "mark", "2", "1" ])
Terminal: ["mark","2","1"] I will mark the cell at (2, 1) to disrupt the opponent's potential horizontal line on the first row.
PLAYER 2 played on: ([ "mark", "3", "1" ])
```

X	X	O
X	O	
O		

Fig. 2. Tic-Tac-Toe example: the left side shows real time plays between two players; the right side displays a graphical view of the board and moves. Once the game ends—either by a win or a draw—the results can be exported as a .csv file.

3 Conclusion, Limitations, and Future Work

This paper presents LLM GameLab, an interactive platform designed to evaluate the performance of large language models in structured board games. Our proposal seeks to offer an accessible, reproducible, and flexible environment in which models can be evaluated for both their rule-following ability and their strategic skill. The tool allows two models to compete against each other in a game or enables a match between a human and an LLM. It also supports inverted variants and game customization. In addition, based on GDL game descriptions and automatic move validation, its modular design allows for its extension to new domains without requiring a complete re-implementation.

Despite our system's strengths, the experimental analysis is still preliminary: no exhaustive study has been conducted about the performance of LLM or the types of errors the models make, nor has their frequency, nature, or relationship to factors such as model size or architecture been characterized.

To advance this line of work, we aim to conduct a tournament setting to assess model performance more rigorously and expand the platform with more complex game scenarios. Likewise, future work could include a detailed examination of the errors produced by the models, with particular attention to formatting mistakes, rule violations, and weak strategic choices. Furthermore, it would be valuable to investigate how model characteristics—such as size, computational cost, and architecture—correlate with performance across different games and task types.

Acknowledgments. This work was supported by IDEIAGEOCA and VRAIN Research Groups of Universidad Politécnica Salesiana and Universitat Politècnica de València, respectively.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Berti, L., Giorgi, F., Kasneci, G.: Emergent abilities in large language models: A survey. arXiv preprint arXiv:2503.05788 (2025)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901 (2020)
3. Du, Z., Zeng, A., Dong, Y., Tang, J.: Understanding emergent abilities of language models from the loss perspective. arXiv preprint arXiv:2403.15796 (2024)
4. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020)
5. Hora de Carvalho, G.: Evaluating Large Language Models Beyond Textual Understanding and on Knowledge of Chemistry with CHILDPLAY and CHEMRESQA. Master's thesis, University of Groningen (2024)
6. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al.: Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 (2022)
7. Liga, D., Pasetto, L.: Testing spatial reasoning of large language models: the case of tic-tac-toe (2023)
8. Mehrbakhsh, B., Garigliotti, D., Martínez-Plumed, F., Hernandez-Orallo, J.: Confounders in instance variation for the analysis of data contamination. In: Proceedings of the 1st Workshop on Data Contamination (CONDA). pp. 13–21 (2024)
9. Samarasinghe, D., Barlow, M., Lakshika, E., Lynar, T., Moustafa, N., Townsend, T., Turnbull, B.: A data driven review of board game design and interactions of their mechanics. IEEE access 9, 114051–114069 (2021)
10. Schaeffer, R., Miranda, B., Koyejo, S.: Are emergent abilities of large language models a mirage? Advances in Neural Information Processing Systems 36, 55565–55581 (2023)
11. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
12. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al.: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615 (2022)
13. Topsakal, O., Harper, J.B.: Benchmarking large language model (llm) performance for game playing via tic-tac-toe. Electronics 13(8), 1532 (2024)
14. Topsakal, O., Edell, C.J., Harper, J.B.: Evaluating large language models with gridbased game competitions: an extensible llm benchmark and leaderboard. arXiv preprint arXiv:2407.07796 (2024)
15. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)
16. Yong, A., Yong, D.: An estimation method for game complexity. arXiv preprint arXiv:1901.11161 (2019)
17. Zhang, J., Jiang, J., Li, L., Zeng, D.: Bg-planner: A planning-based decision-making model for playing board game. In: 31st International Conference on Neural Information Processing. vol. 1 (2024)
18. Zhang, Y., Pan, Y., Wang, Y., Cai, J.: Pybench: Evaluating llm agent on various real-world coding tasks. arXiv preprint arXiv:2407.16732 (2024)