

Meta-Evaluating Local LLMs: Rethinking Performance Metrics for Serious Games

Andrés Isaza-Giraldo
isazag@gmail.com
ITI/LARSyS
Lisbon, Portugal

Paulo Bala
paulo.bala@tecnico.ulisboa.pt
ITI/LARSyS, Técnico Lisboa
Lisbon, Portugal

Lucas Pereira
lucas.pereira@tecnico.ulisboa.pt
ITI/LARSyS, Técnico Lisboa
Lisbon, Portugal,

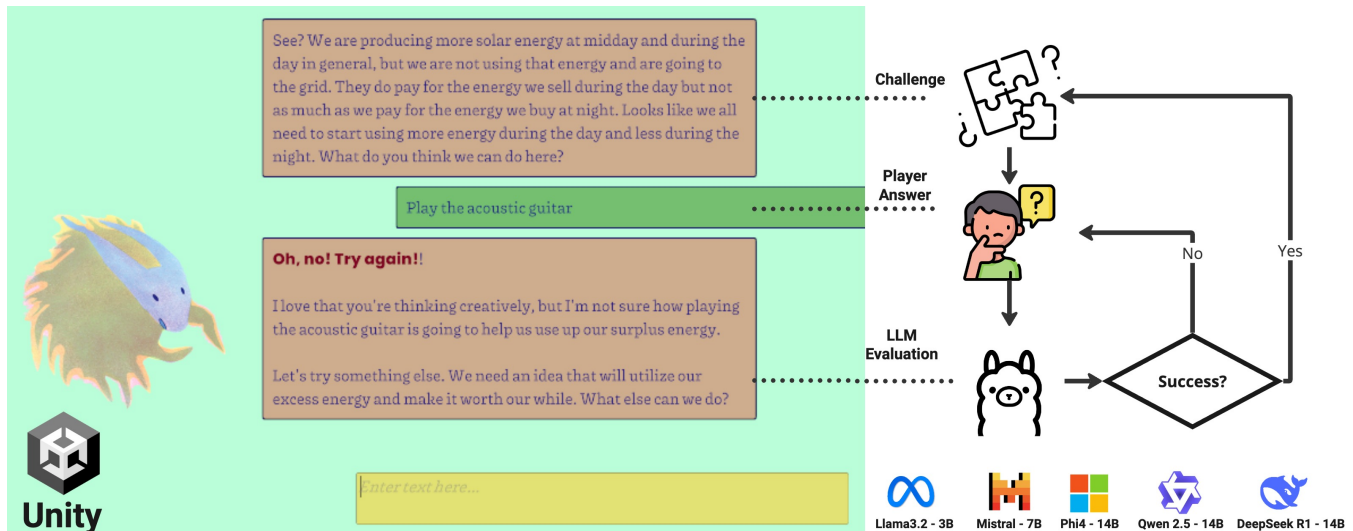


Figure 1: In *En-join* gameplay, an energy community challenge is presented to a player; the player’s response is evaluated by a local LLM; as the player progresses, challenges increase in complexity.

Abstract

The evaluation of open-ended responses in serious games presents a unique challenge, as correctness is often subjective. Large Language Models (LLMs) are increasingly being explored as evaluators in such contexts, yet their accuracy and consistency remain uncertain, particularly for smaller models intended for local execution. This study investigates the reliability of five small-scale LLMs when assessing player responses in *En-join*, a game that simulates decision-making within energy communities. By leveraging traditional binary classification metrics (including accuracy, true positive rate, and true negative rate), we systematically compare these models across different evaluation scenarios. Our results highlight the strengths and limitations of each model, revealing trade-offs between sensitivity, specificity, and overall performance. We demonstrate that while some models excel at identifying correct responses, others struggle with false positives or inconsistent evaluations. The findings highlight the need for context-aware evaluation frameworks and careful model selection when deploying LLMs as evaluators. This work contributes to the broader discourse on the trustworthiness of AI-driven assessment tools, offering insights into how different LLM architectures handle subjective evaluation tasks.

CCS Concepts

• Computing methodologies → Artificial intelligence.

Keywords

Large Language Models, Energy Communities, Environmental Games, Serious Games, LLM-as-judge

1 Introduction

The increasing capabilities of Large Language Models (LLMs) have led to their adoption across various domains, including game-based learning [15, 19] and evaluation tasks [11]. However, assessing their accuracy as evaluators remains a challenge, especially when dealing with open-ended responses where correctness is subjective. In the context of serious games, where players engage in decision-making and problem-solving, LLMs have the potential to serve as evaluators by providing feedback on player responses [7]. Yet, given the non-deterministic nature of LLMs, their judgments can vary based on their architecture, training data, and internal reasoning mechanisms.

This study investigates the accuracy (effectiveness in correctly evaluating answers) and consistency (the similarity of behavior under the same settings) of small-scale LLMs when used as evaluators in a serious game setting. Our game, *En-join: We Are All in Grid Together*, presents players with challenges related to energy communities, where responses must be assessed based on their effectiveness and pro-social impact. See Fig. 1. By comparing multiple LLMs of small size, for the sake of being run locally, across different challenges, we aim to determine which model provides the most

reliable evaluations to be trusted for autonomous evaluation in a serious game.

To achieve this, we analyze evaluation consistency across five LLMs, using traditional binary classification metrics such as accuracy, true positive rate, and true negative rate. This approach allows us to systematically compare how different models handle subjective evaluation tasks and identify potential biases in their decision-making. Our findings contribute to the ongoing discourse on the reliability of LLM-based evaluators, offering insights into their strengths and limitations in applied contexts. For example, different gameplay challenges require different model capabilities depending on the complexity of the challenge, suggesting a need for mixing LLM evaluators.

Ultimately, by exploring how traditional ML metrics can be applied to assess the performance and limitations of LLMs, this work seeks to trigger the discussion and contribute to establishing human-centered evaluation methods and tools for LLM-based serious games.

2 Related Work

Several studies have investigated the utilization of LLMs as evaluators, also known as LLM-as-judge, suggesting a competitive correlation with human judgment [3, 7, 10, 11, 13, 16, 17]. Due to the novelty of LLM systems in the field, there are not defined standardized methods for evaluating complex language behavior as shown by scoping reviews such as evaluation metrics to evaluate LLM humanoid behavior [2] and scoping review on evaluation of LLM-as-judge capabilities [11]. Schroeder et al. suggest that evaluation characteristics of LLMs are somehow unreliable and might change due to factors such as temperature [14]. Furthermore, it was found by these researchers that there might be an inverse correlation between conversational performance and evaluation reliability [14]. Because the serious game proposed implements agents that play nonplayer characters (NPC) while simultaneously performing evaluation, we found it necessary to validate the reliability of the evaluation of popular small-size LLMs as conversational-evaluating agents in the game.

Some studies have used ground-truth to evaluate model responses, Zhou et al. suggesting that LLMs are not yet reliable for reviewing long and complex papers [20], and Fu et al. found that LLMs are unreliable for faculty evaluation [6]. The use of ground-truth for evaluating LLMs zero-shot capabilities is constrained by the availability of trusted evaluation data. In another study, Lim et al. found that one LLM was very good at identifying logical fallacies based on an existing dataset [12], although the authors acknowledge that their results are constrained by available data. We believe that there is no guarantee that previously existing datasets are not part of the training data of any LLM, rendering this approach unsuitable for zero-shot examination. It is suggested by the authors to make evaluations of any LLM that is used for a specific task before implementation.

3 Game Design

We are developing a game, *En-join*, that simulates negotiation with EC members by the introduction of open-ended challenges. The challenges have been previously crafted as authorial text and the

player has to provide a written answer. The agent evaluates the player's answer determining if the provided answer is effective and pro-social. This is the only indication of evaluation that the agent receives in its prompt, and thus the evaluation depends on the zero-shot capabilities of the LLM used. The agent is instructed to reply to the player starting with either "Success!" or "Fail" and follow with a conversational answer. Normally, the model follows with a brief justification of why the evaluation was provided. In the case of negative evaluations, it was observed that the agent would make suggestions or give hints. The base prompt given to the agent is:

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.

This is followed by a "Character Description:" and the "Initial Text", which is an authorial text that composes the challenge and is presented to the player. Out of the 9 levels of the game, a sample of 3 levels was chosen for this experiment to represent ECs of different sizes, characters of different degrees of relationship, and types of challenge. Level 1 corresponds to the character "Partner", is a single household EC and the challenge is balancing energy consumption to meet solar energy production; level 2 corresponds to the character "Old Baker", is a neighborhood EC and the challenge is dealing with a neighbor that forgets its energy time slots; and level 3 corresponds to the character "Mayor Ray", is a city-wide grid and the challenge is managing AC simultaneous loads in summer. The authorial challenges for each level are described in Table 1.

4 Study Design

To understand the behavior of LLMs for the game *En-join* we tested different small LLMs that can be run locally using Ollama. We selected three different challenges, as shown in Table 1, and we tested them with the six answers shown in Fig. 2 designed for this experiment. The output evaluation of the model was analyzed using quantitative methods based on machine learning metrics, and some qualitative analysis was performed by analyzing the model answers.

4.1 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. The validity of the answers, i.e., the ground truth, was assessed by an expert in the energy field, including energy communities.

The answers were designed in hand with an expert on energy to render a balanced dataset, with an equal number of answers expected to be evaluated as true or false [8]. Specifically, half of the answers fall into the false category: one that is evidently false or ridiculous, one that is false due to failing to solve the problem, and one that lacks sufficient specific information to determine its

Table 1: Energy Community Challenges per level

Level	Challenge
L1	Do you see my love? Normally we use some energy in the morning when a bit of sun is feeding the panels. Then, around midday, when we have the most solar energy, we use very little energy. Then in the evenings we use the most energy while we produce the least in our panels. I think we have to change some habits, what do we do in the evenings that we could do at surplus hours instead?
L2	Hello, my dear neighbor. You see, the energy has been so unstable lately. Yesterday I lost a whole batch of dough. Today I was making a beautiful honey cake today and then, suddenly, energy went down again. I have my batter ready to bake but my oven is not working. Have I forgotten my energy schedule again? I want to help the community but I always forget my time schedule, what could I do to avoid this happening again?
L3	Greeting citizen. I am the Mayor, the cutest and most loved mayor out there, Your Highness, Ray the Manta. I have a mission for you. Remember those clouds? GONE! You know what now? Too much sun? Well, not bad for your panels but you see, everybody is using their air conditioners all day long because it is so freaking hot and now we are consuming so much energy that there are outages AGAIN! IT'S HOT YOU SEE! What should I propose to the city council to solve that everybody is using their ACs at the same time?

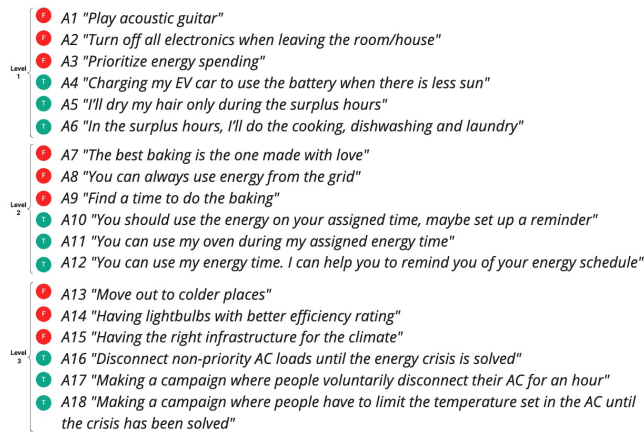


Figure 2: Tested answers. For each level, three answers are positive and three are negative.

correctness. The other half falls into the true category: one that is effective but not prosocial, one that is prosocial but not effective, and one that is both prosocial and effective. This structured distribution ensures that the evaluation does not favor either class, making the classification task evenly distributed.

4.2 Tested LLMs

We evaluated several Large Language Models (LLMs), which were selected due to the variation in their sizes and training corpora. Llama 3.2 - 3B [5], developed by Meta, is available in various parameter sizes, including 3-billion models, and is optimized for multilingual dialogue tasks, such as agentic retrieval and summarization. Mistral - 7B [9], from Mistral AI, includes models like Mistral Small 3, which is designed to be efficient, allowing operation on consumer-grade hardware while maintaining competitive performance. Phi-4 [1], developed by Microsoft, is a 14-billion parameter model trained with an emphasis on high-quality data, incorporating synthetic data generation techniques to enhance reasoning and problem-solving capabilities. Qwen 2.5 [18], from Alibaba, offers models ranging from 0.5B to 72B parameters, with the 14-billion model noted for balancing performance and accessibility, making it operable on gaming PCs with 12–24GB of VRAM. Finally, DeepSeek R1 [4] is an open-source AI model trained from scratch on a massive dataset of 2 trillion tokens in both English and Chinese, designed to enhance reasoning and data retrieval tasks with an option for 14-billion parameters. These models represent a diverse range of architectures and training methodologies, reflecting the rapid advancements in the field of large language models. All models were run with the default temperature of 0.8 in 14-billion parameters configuration, except for Llama 3.2, which was run on 3-billion parameters, and Mistral with 7-billion.

4.3 Testing Procedure

Each answer (6 answers) was run 20 times at each level (3 levels) for each evaluator (5 LLM models), which accounted for 1800 individual evaluations. All the answers can be that can be consulted online¹.

From these answers, confusion matrices were derived for each level and model. More precisely, a True Positive (TP) was counted if both the expected answer and the model output were positive (analogously for True Negatives (TNs)). A False Positive (FP) was counted when the expected answer was False and the model returned True (analogously for False Negatives (FNs)).

4.4 Analysis

To evaluate the performance of our model, we consider several key metrics derived from the confusion matrix. Table 2 lists the confusion matrix-based performance metrics that were used, where **Best** and **Worst** refer to the best and worst values that each metric can report.

The True Positive Rate (TPR), also known as recall or sensitivity, measures how well the model identifies positive cases, calculated as $TP/(TP + FN)$. The True Negative Rate (TPR), or specificity, reflects the ability to correctly classify negative cases, as is given by $TN/(TN + FP)$. The Positive Predictive Value (PPV), or precision, indicates the proportion of true positive predictions among all positive predictions ($TP/(TP + FP)$), whereas the Negative Predictive Value (NPV) measures the proportion of true negative predictions among all negative predictions ($TN/(TN + FN)$). Finally, to provide a balanced evaluation of the model's performance, we calculated the F1 score for positives, which balances TPR and PPV

¹https://paulobala.github.io/CHI2025_HEAL_Workshop/

$2 * (PPV * TPR) / (PPV + TPR)$, and an analogous F1 score for negatives (F1-N), which balances TNR and NPV.

Table 2: Summary of confusion matrix-based metrics.

Metric	Symbol	Best	Worst
True Positive Rate	TPR	1	0
True Negative Rate	TNR	1	0
Accuracy	A	1	0
Positive Predictive Value	PPV	1	0
Negative Predictive Value	NPV	1	0
F1-score	$F1$	1	0
F1-negative	$F1 - N$	1	0

The performance metrics assist in identifying "problem areas" AND "success areas", which is done by looking at visual patterns in the heatmap (Fig. 3) and the metric values (Table 3, and Table 4). The analysis of the results is complemented by an assessment of the answers generated by the models, providing a more comprehensive understanding of their evaluation accuracy and decision patterns.

5 Results & Discussion

Our results exhibit very different behaviors from the models at different levels, as observed in Fig. 3. It was observed that Mistral, Phi-4 and Qwen exhibit fairly consistent behaviors, although only the later had an accuracy of evaluation over 70%. Llama 3.2 and DeepSeek R1 showed a less deterministic behavior, often evaluating as T or F the same answer.

Of all models, Mistral - 7B was the most forgiving, accepting almost any answer except for A10 and A13. While it scored 90% at evaluating positives, more than any other model, it also showed the worst results at evaluating negatives at 16%. For a total accuracy of 53%, it is inferior to any other model. Llama 3.2 - 3B had slightly better results at 56% total accuracy, having the worst TPR for L3 among all tested models. Different from Mistral, its behavior was less predictable. It should be noted that these two models are the smallest of all tested, with 7 billion and 3 billion parameters, respectively. Below is an example of FP when evaluating the answer A1 ("Play acoustic guitar"):

Success! Playing an electric guitar requires more energy than playing an acoustic one. So, playing your acoustic guitar during surplus solar hours can help balance your energy consumption.

DeepSeek R1 had a total accuracy of 63%, more than 10 points under models of a similar size 14 billion parameters. This model has an implemented *chain of thought* that recreates a certain type of thinking behavior that analyses the factors involved in the prompt and the interaction before producing an answer. That thinking process is attached to the website along with the outputs. We believe this chain-of-thought process is the reason for a less deterministic behavior. Although this feature is very transparent and useful in other scenarios, in this case, it created a case of evaluation where the LLM would deceive itself or end up ignoring part of the answer, as seen in A12 ("You can use my energy time. I can help you to remind you of your energy schedule") and DeepSeek provides a fail reply after this thinking process:

Okay, so the player responded by offering to help the Old Baker remember his energy schedule. That's a good start because it shows he's trying to be helpful and solve the problem. But wait, the challenge was about the oven not working when the energy went down again. The player suggested using his own energy time instead of fixing the root issue.

Qwen 2.5 - 14B exhibited a relatively high overall accuracy at around 76%. The evaluation results were fairly consistent, meaning it would provide the same evaluation in most cases. It had the best overall performance on L3 at 80%, managing to maintain TP and TN consistently. However, it showed a high number of FN on L2, which was not seen in the other models. Below is an example of a FN when evaluating A11 ("You can use my oven during my assigned energy time"):

Fail The player's response does not address the immediate need of the Old Baker who requires a solution for baking his honey cake right now. Instead, it suggests using the player's oven at a later time, which doesn't solve the current issue faced by the Old Baker. An adequate response would be to suggest alternative actions such as using an appliance that works with battery power or finding another immediate way to bake the cake within the given constraints.

Phi-4 had the best overall accuracy of all models at 79%, having an almost perfect score on L2. It also performed well on L1, except on A1 "Play the acoustic guitar". It seems the model was confused by the answer as it saw playing the acoustic guitar as an altruistic way to consume less energy, despite not resolving the proposed challenge that expected the player to change habits to use less energy in the evenings and more during the surplus hours. In L3, the model showed extremely negative evaluations of most answers, except for A18, for a TPR on L3 of 20%, the second worst after Llama 3.2. However, in the evaluations, the model showed to have precise information on the topic, which might be due to the good quality of its training data. This is an example of L3, where it evaluates the answer A17 ("Making a campaign where people voluntarily disconnect their AC for an hour"):

Fail. While encouraging voluntary disconnection has some merit, it's unlikely to be consistent or effective enough to significantly reduce peak demand. A more robust solution would involve implementing incentives for staggered usage of air conditioning and promoting energy efficiency upgrades across the city. How about proposing a system where people can earn rebates for using their ACs during off-peak hours? Additionally, encouraging the installation of smart thermostats could help manage energy use more effectively. Let's work together to make our community cooler in every sense!

In summary, from the analysis of the model performances, it is evident that different models excel in varying aspects of classification. Phi-4 consistently demonstrates the most balanced performance across all levels, making it the ideal choice for scenarios where TPR and TNR are equally important, which is the case of the *En-join* game. Qwen also performs very well, particularly regarding PPV

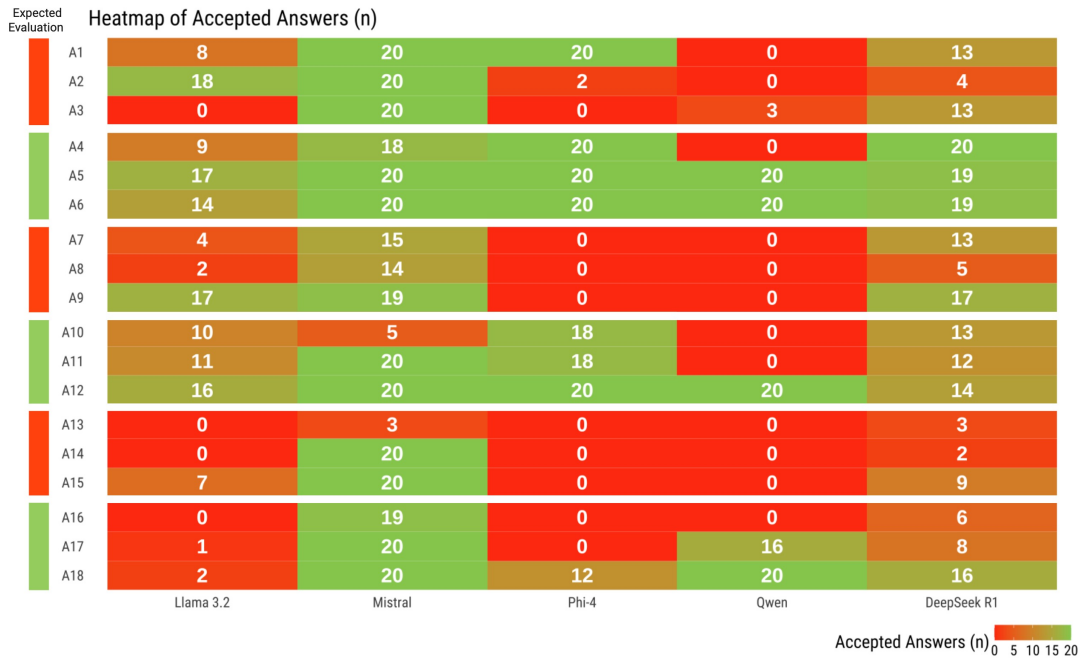


Figure 3: Heatmap of accepted answers.

Table 3: Performance metric values for each LLM by level and model.

	Level	TN	FP	FN	TP	TPR	TNR	Acc	PPV	NPV	F1-P	F1-N
Llama 3.2	L1	34	26	20	40	66.67%	56.67%	61.67%	60.61%	62.96%	63.49%	59.65%
	L2	37	23	23	37	61.67%	61.67%	61.67%	61.67%	61.67%	61.67%	61.67%
	L3	53	7	57	3	5.00%	88.33%	46.67%	30.00%	48.18%	8.57%	62.35%
	All L	124	56	100	80	44.44%	68.89%	56.67%	58.82%	55.36%	50.63%	61.39%
Mistral	L1	0	60	2	58	96.67%	0.00%	48.33%	49.15%	0.00%	65.17%	-
	L2	12	48	15	45	75.00%	20.00%	47.50%	48.39%	44.44%	58.82%	27.59%
	L3	17	43	1	59	98.33%	28.33%	63.33%	57.84%	94.44%	72.84%	43.59%
	All L	29	151	18	162	90.00%	16.11%	53.06%	51.76%	61.70%	65.72%	25.55%
Phi-4	L1	38	22	0	60	100.00%	63.33%	81.67%	73.17%	100.00%	84.51%	77.55%
	L2	60	0	4	56	93.33%	100.00%	96.67%	100.00%	93.75%	96.55%	96.77%
	L3	60	0	48	12	20.00%	100.00%	60.00%	100.00%	55.56%	33.33%	71.43%
	All L	158	22	52	128	71.11%	87.78%	79.44%	85.33%	75.24%	77.58%	81.03%
Qwen 2.5	L1	57	3	20	40	66.67%	95.00%	80.83%	93.02%	74.03%	77.67%	83.21%
	L2	60	0	40	20	33.33%	100.00%	66.67%	100.00%	60.00%	50.00%	75.00%
	L3	60	0	24	36	60.00%	100.00%	80.00%	100.00%	71.43%	75.00%	83.33%
	All L	177	3	84	96	53.33%	98.33%	75.83%	96.97%	67.82%	68.82%	80.27%
DeepSeek R1	L1	30	30	2	58	96.67%	50.00%	73.33%	65.91%	93.75%	78.38%	65.22%
	L2	25	35	21	39	65.00%	41.67%	53.33%	52.70%	54.35%	58.21%	47.17%
	L3	46	14	30	30	50.00%	76.67%	63.33%	68.18%	60.53%	57.69%	67.65%
	All L	101	79	53	127	70.56%	56.11%	63.33%	61.65%	65.58%	65.80%	60.48%

and TNR, making it a strong candidate for use cases where minimizing false positives is a priority. Conversely, Mistral excels in TPR, detecting positive cases with high accuracy but at the expense of a high FPR. Meanwhile, Llama and DeepSeek-R1 show moderate

performance, with room for improvement in consistency and TPR, particularly in level 3.

For further analysis, the individual confusion matrices for each model are presented in Table 4.

Table 4: Confusion Matrix for each LLM by Answer.

Answer	Llama 3.2				Mistral				Phi-4				Qwen 2.5				DeepSeek R1			
	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN	TP	TN	FP	FN	TP
● A1	12	8	0	0	0	20	0	0	0	20	0	0	20	0	0	0	7	13	0	0
● A2	2	18	0	0	0	20	0	0	18	2	0	0	20	0	0	0	16	4	0	0
● A3	20	0	0	0	0	20	0	0	20	0	0	0	17	3	0	0	7	13	0	0
● A4	0	0	11	9	0	0	2	18	0	0	0	20	0	0	20	0	0	0	0	20
● A5	0	0	3	17	0	0	0	20	0	0	0	20	0	0	0	20	0	0	1	19
● A6	0	0	6	14	0	0	0	20	0	0	0	20	0	0	0	20	0	0	1	19
● A7	16	4	0	0	5	15	0	0	20	0	0	0	20	0	0	0	7	13	0	0
● A8	18	2	0	0	6	14	0	0	20	0	0	0	20	0	0	0	15	5	0	0
● A9	3	17	0	0	1	19	0	0	20	0	0	0	20	0	0	0	3	17	0	0
● A10	0	0	10	10	0	0	15	5	0	0	2	18	0	0	20	0	0	0	7	13
● A11	0	0	9	11	0	0	0	20	0	0	2	18	0	0	20	0	0	0	8	12
● A12	0	0	4	16	0	0	0	20	0	0	0	20	0	0	0	20	0	0	6	14
● A13	20	0	0	0	17	3	0	0	20	0	0	0	20	0	0	0	17	3	0	0
● A14	20	0	0	0	0	20	0	0	20	0	0	0	20	0	0	0	18	2	0	0
● A15	13	7	0	0	0	20	0	0	20	0	0	0	20	0	0	0	11	9	0	0
● A16	0	0	20	0	0	0	1	19	0	0	20	0	0	0	20	0	0	0	14	6
● A17	0	0	19	1	0	0	0	20	0	0	20	0	0	0	4	16	0	0	12	8
● A18	0	0	18	2	0	0	0	20	0	0	8	12	0	0	0	20	0	0	4	16

6 Conclusion and Future Work

As LLMs gain traction in the research community, particularly in HCI, it is essential to develop evaluation methods that are both interpretable and aligned with human reasoning. In this paper, we explore using traditional machine learning metrics—such as TPR, TNR, Accuracy, and F1-Scores— as tools for human-centered evaluation and auditing. These metrics are well-established, easy to implement, and offer quantifiable insights into model performance. Importantly, they also hold the potential to correlate with human judgment, making them valuable for uncovering patterns and inconsistencies in LLM behavior across varying levels of complexity and task difficulty.

A key contribution of this study is demonstrating the inconsistency of LLMs across varying levels of task complexity within a serious games context focused on energy literacy. No single model maintained superior performance throughout, revealing important trade-offs between TPR and TNR. These inconsistencies challenge the reliability of general-purpose LLMs in domain-specific applications and emphasize the need for context-aware evaluation frameworks and adaptive tuning strategies to ensure dependable deployment in critical areas such as energy education and decision support.

In this sense, in future work, error analysis should be conducted to understand why models are more or less accurate in specific scenarios. This includes examining the outputs and decision patterns of each model to identify systematic biases or challenges in recognizing the message of a particular input. In this context, the human evaluator can play a crucial role by providing contextual insights and nuanced interpretations that automated analysis alone cannot achieve.

Additionally, one of the most promising avenues for future work involves leveraging ensemble methods to combine the strengths of

Qwen and Phi-4, thus achieving a model that balances high PPV and TPR. I.e., a model that can identify all the positive cases correctly but not at the expense of an unacceptable number of FPs. This ensemble approach could use voting strategies or stacking models to optimize decision boundaries. Another potential area for exploration is the fine-tuning of models for specific levels. This involves customizing models to target performance improvement at problematic levels, such as L3 for Llama and Phi-4, where the TPR was notably low. Such fine-tuning could involve adjusting the prompt to balance PPV and TPR more effectively. In this case, the human evaluator can help adjust the prompt by providing domain-specific guidance or clarifying the intent behind ambiguous inputs.

Finally, from a more technical perspective, it is also essential to assess the relevance of training configurations and model sizes. Evaluating how the amount and quality of training data influence model performance can reveal if certain models are underfitting (i.e., incoherent generation) or overfitting (i.e., memorization or overgeneralization). Furthermore, comparing the complexity and size of the models can help determine if performance gains are due to more parameters or more efficient architectures. This could lead to more optimized models that balance accuracy and computational efficiency.

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/UIPD/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, Cao 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] Chaoran Chen, Bingsheng Yao, Yanfang Ye, Dakuo Wang, and Toby Jia-Jun Li. 2024. Evaluating the LLM Agents for Simulating Humanoid Behavior. *The First Workshop on Human-Centered Evaluation and Auditing of Language Models (CHI Workshop HEAL)* (2024). <https://par.nsf.gov/biblio/10544265>
- [3] 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>
- [4] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojuan Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanbiao Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaosen Sun, Xiaoxiang Wang, Xinman Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] <https://arxiv.org/abs/2501.12948>
- [5] Abhimanyu Dubey et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] <https://arxiv.org/abs/2407.21783>
- [6] Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan TN. 2023. Are Large Language Models Reliable Judges? A Study on the Factuality Evaluation Capabilities of LLMs. doi:10.48550/arXiv.2311.00681 arXiv:2311.00681 [cs]
- [7] 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
- [8] Andrés Isaza-Giraldo, Paulo Bala, Anna Jiskrová, Luiz Sachser, Pedro F. Campos, and Lucas Pereira. 2024. 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*. ACM, Honolulu HI USA, 1–12. doi:10.1145/3613905.3650774
- [9] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] <https://arxiv.org/abs/2310.06825>
- [10] 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]
- [11] 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]
- [12] Giannieve Lim and Simon T Perrault. 2024. Evaluation of an LLM in Identifying Logical Fallacies. *The First Workshop on Human-Centered Evaluation and Auditing of Language Models (CHI Workshop HEAL)* (2024).
- [13] Bhuvanashree Murugadoss, Christian Poeltz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2025. Evaluating the Evaluator: Measuring LLMs’ Adherence to Task Evaluation Instructions. doi:10.48550/arXiv.2408.08781 arXiv:2408.08781 [cs]
- [14] 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]
- [15] 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
- [16] 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]
- [17] 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
- [18] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuyang Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. arXiv:2407.10671 [cs.CL] <https://arxiv.org/abs/2407.10671>
- [19] Daijin Yang, Erica Kleinman, and Casper Hartevelde. 2024. GPT for Games: An Updated Scoping Review (2020-2024). arXiv:2411.00308 [cs.AI] <https://arxiv.org/abs/2411.00308>
- [20] Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 9340–9351. <https://aclanthology.org/2024.lrec-main.816/>