Exploring the use of open-source LLMs for Game Learning Analytics: an empirical study

Defala Filho[1], Fabrizio Honda[1], Fernanda Pires[1], Marcela Pessoa[1]

[1]School of Technology – State University of Amazonas (EST-UEA)
ThinkTEd Lab – Research, Development, and Innovation in Emerging Technologies

{dadff.lic23, fpires, msppessoa}@uea.edu.br

{fabrizio.honda}@icomp.ufam.edu.br

Abstract. *Game Learning Analytics (GLA) is essential for analyzing player behavior and identifying their learning progress through data collection and analysis. However, data modeling for GLA is a complex and challenging process. Large Language Models (LLMs) can be an alternative, especially open-source models with no usage limitations. This study investigates how the models "Gemma 2 (9B)," "Qwen 2.5 (14B)," "LLaMA 3.1 Instruct (8B)," and "Phi-4 Mini (3.8B)" are capable of modeling data and filling out the data capture template from the GLBoard model. Seven GLA experts evaluated each model's responses and identified the "Qwen" model as the most satisfactory, highlighting the potential of open-source LLMs in GLA activities.*

1. Introduction

The assessment of learning in educational games is a fundamental process, making it possible to verify if the player is developing the expected cognitive skills. One of the most common ways to assess these objects is through questionnaires, especially self-assessment [Din et al. 2023], which only considers the students' perception, disregarding the game as an assessment instrument. Thus, educators have no control over what is happening during the students' gameplay [Alonso-Fernández et al. 2021], because the learning traits are correlated with the mechanics' elements [Melo et al. 2020, Van Eck 2006].

Game Learning Analytics (GLA) emerges as an alternative, dealing with collecting, analyzing, and visualizing data from serious games [Freire et al. 2016]. This alternative allows the capture of player interaction records (logs) in a non-intrusive way, avoiding disrupting the game flow. Through GLA data, developers can validate game design, teachers can understand the learning flow and apply interventions, and students can monitor their progress [Banihashem et al. 2024]. Furthermore, by crossing GLA data with heuristic evaluations and data science techniques, we can obtain more valuable insights about the students' learning [Alonso-Fernández et al. 2021, Silva et al. 2021, Alonso-Fernández et al. 2022].

Despite the advantages of GLA and the existence of models that enable the implementation of its techniques, data modeling is a complex and challenging process [Honda et al. 2025b]. This process is a fundamental step in GLA and must be carried out from the beginning of the game's design, referring to the process of defining which data we will collect (GLA variables) and their importance in expressing the player's learning evolution [Hauge et al. 2014, Alonso-Fernández et al. 2021, Kitto et al. 2020]. One can

DOI: 10.5753/sbie.2025.12885

fill out the data template (GLA capture structure) from this step. The difficulties arising from this process are: (i) understanding the game clearly enough to define the capture variables; (ii) abstraction; (iii) time dedicated to this process; and (iv) the need to understand concepts from computer science disciplines (systems modeling, databases, and programming) that facilitate data modeling [Honda et al. 2025b].

An emerging field that can help in this context is the use of Large Language Models (LLMs): models capable of generating texts like humans and performing various tasks [Kasneci et al. 2023]. Interaction with these models usually occurs via a chatbot, where the user sends instructions (a prompt) and receives a response from the LLM, artificially generated from its knowledge base. Researchers have applied LLMs to various domains, such as Serious Games [Mitsea et al. 2025] and Learning Analytics [Misiejuk et al. 2025]. Furthermore, we have recently observed their use in GLA contexts, which have presented a limitation of closed models (such as ChatGPT, whose model source code is not available for training) [Honda et al. 2024].

Considering (i) the contributions of the GLA field, (ii) the use of LLMs for diverse domains, such as GLA, and (iii) the limitations of closed models in some aspects, this work presents as a research question (RQ): "how do open-source LLMs perform in GLA activities, especially in filling data templates?".

2. Foundations and related work

GLA involves collecting, capturing, and analyzing data from serious games to track players' progress through levels. This strategy allows for the analysis of player behavior, the identification of evidence of learning, and the creation of profiles, among other tasks. Among the tools that enable the implementation of GLA techniques is GLBoard: a model to assist in collecting and visualizing data from educational games [Silva et al. 2022]. It consists of four modules: (i) Unity Package, which allows incorporating the model into games, providing a JSON model for data collection; (ii) Database, which stores player interaction records and game information; (iii) API, which manages the system and communicates with other modules; and (iv) Dashboard, which allows for the visualization of raw data or visual analysis through graphs. After incorporating GLBoard into the game, the developer can start filling in data in the model, which contains four classes: $Player\_Data$ (with player information), $Game\_Data$ (game-related information), $Phase$ (the available phases), and $Section$ (records the game sections, i.e., each player's attempt).

The template contains predefined variables; the developer must fill them in via code. The variable that presents a different behavior is $path\_player$, corresponding to the player's path through the levels. It allows adding any GLA variable, as it is a "List" object of type "string", making GLBoard flexible for any educational game. However, it is necessary to define the GLA data and the justification for why they are essential to identify the player's learning progression. This step is data modeling, which should occur from the beginning of the game's creation [Hauge et al. 2014]. When a game is already partially implemented or finalized, including the data template becomes challenging even for developers with GLA experience. This challenge occurs because the learning designers did not model the data, causing difficulties in incorporating the GLA template, such as game abstraction, complex mechanics, structure adaptation, and navigating the code [Macena et al. 2024].

Recently, we've noticed that researchers are applying LLMs in the context of GLA, particularly in the data modeling process [Honda et al. 2024]. However, research has highlighted limitations of closed-source models (such as ChatGPT), particularly regarding usage restrictions, payment requirements, and other factors. In this regard, open-source models may be a viable alternative. We can download these models from platforms such as HuggingFace for training on domain-specific datasets (fine-tuning), facilitating customization. There are a variety of open-source models, whose researchers generally evaluate them based on benchmarks. These benchmarks consist of metrics and standardized tasks that allow comparing models, such as accuracy, consistency, adaptability, etc. For this study, we sought research that addresses at least two of the following: LLMs (open or otherwise), Learning Analytics (LA), Serious Games, and data modeling, described below.

Alonso-Fernandez et al. [2019] showed how using Learning Analytics can improve the development and application of serious games in educational settings. The authors analyzed three games: Conectado (bullying), DownTown (subway use by people with intellectual disabilities), and First Aid Game (cardiopulmonary resuscitation for adolescents). They collected player data via GLA and used it to provide feedback. The authors conclude that GLA is essential for the pedagogical validation of games, with tests conducted with more than 300 students between the ages of 12 and 17, highlighting the importance of planning data collection from the game's conception.

Kitto et al. [2020] proposed approaches for generating relevant educational data aligned with pedagogical objectives in Learning Analytics ecosystems. Through a conceptual and analytical study, the authors analyzed common limitations in educational data collection without participant testing. They advocated a top-down approach, starting from learning needs to structure data extraction. Using examples such as the BeyondLMS project, the study reinforces that data quality and usefulness depend on planning guided by educational constructs and collaboration between educators and developers.

Honda et al. [2024] developed a conversational agent in GLA to assist developers in data modeling in educational games using GLBoard. The authors built an application in ChatGPT's "MyGPTs" tool, which can generate data capture structures in JSON (GLBoard template) from user interactions. The research adopted the Design-Based Research (DBR) methodology, with iterative development and validation cycles. The authors composed the agent's knowledge base with GLBoard technical documentation, relevant articles in the field of GLA, and examples. The evaluation included seven GLA experts in Brazil, who tested the agent and answered quantitative and qualitative questionnaires. The results indicated that the agent generated relevant responses and structures compatible with GLBoard.

Table 1 summarizes the main characteristics of each work, comparing them with the present study. None of the works meets all the criteria. The most common ones focus on using LLMs for LA or incorporating LA in Serious Games. Although Honda et al. [2024] propose an AI agent to assist in data modeling, they focus on closed models without exploring open-source alternatives or using benchmarks. This work's difference is in combining LA, Serious Games, data modeling, and open-source LLMs.

Table 1. Comparison between related works.

Work	LA	Data Modeling	Serious Games	Open-source LLMs
Alonso-Fernández et al. [2019]	X	–	X	–
Kitto et al. [2020]	X	X	–	–
Honda et al. [2024]	X	X	X	–
This work	X	X	X	X

3. Methods

This research aims to analyze how open-source LLMs perform in Game Learning Analytics (GLA) activities, particularly in generating data capture structures compatible with the GLBoard model. Models must perform adequate data modeling to successfully perform this activity, which we also intend to investigate. Therefore, to achieve this objective, we conducted an empirical study that included seven steps, illustrated in Figure 1 and described below.

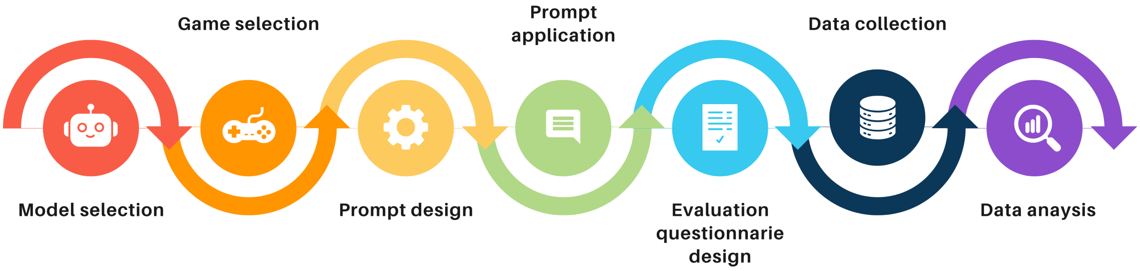


Figure 1. Steps taken in the empirical study.

3.1. Model selection

The first stage refers to the definition of the aspects related to the LLMs we use: (i) the selection of models, (ii) the prompting technique adopted, (iii) the use of context-based learning (ICL), and (iv) the model execution environment.

The selection of models began with the analysis of the Open LLM Leaderboard[1]: a classifier available on the Hugging Face platform that provides a systematic evaluation of the performance of several open-source LLMs, based on standardized benchmarks. Researchers test each model available in this ranking on different benchmarks, which evaluate different competencies. In the context of GLA-related tasks, the most relevant benchmarks are: IFEval (measures the model's ability to follow instructions accurately) [Zhou et al. 2023], BHH – Big-Bench Hard (evaluates logical reasoning, common sense, and complex problem solving) [Suzgun et al. 2022], and MuSR – Multi-Step Reasoning (examines the performance of models in tasks with multiple steps and context dependence).

Therefore, based on the Open LLM Leaderboard, we selected and cataloged the models with the highest scores in these benchmarks in a spreadsheet. We mapped A total

[1] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard\#/

of 12 models, and to determine which ones would we use, we defined the following criteria: (i) number of parameters – we chose models with 3B to 15B parameters due to the limitations of the computational infrastructure for executing the models; (ii) average performance in the benchmarks – prioritizing models with the highest averages in the IFEval, BHH and MuSR benchmarks; and (iii) preference for "original" models – excluding the distilled ones, which inherit behavior patterns from larger models from supervised tuning, as they could compromise the autonomous capacity of the models. Therefore, we selected four models, whose names, parameters, benchmark scores, and averages are available in Table 2.

Table 2. Selected models and their scores on benchmarks.

Model	Parameters	IFEval	BBH	MuSR	Average
phi4-mini	3B	73.78	38.74	6.45	39.65
llama3.1-instruct	8B	49.22	29.38	8.61	29.07
Qwen2.5	14B	36.94	45.08	15.91	32.64
Gemma2	9B	74.36	42.14	9.74	42.14

The prompting technique is related to prompt engineering, which refers to the design of appropriate prompts to obtain the desired response [Lo 2023]. Considering the emergence of the area, although studies seek to systematize them [Liu et al. 2023, Sahoo et al. 2024], the emergence of new prompting techniques and similar terminologies for the same technique, among others, can be noted. In this aspect, the method depends on the context in which we use it. In this study, we chose basic prompting due to its easy understanding and use: generating a simple prompt to the model to obtain the desired results [Shin et al. 2023, Brown et al. 2020, Liu et al. 2023].

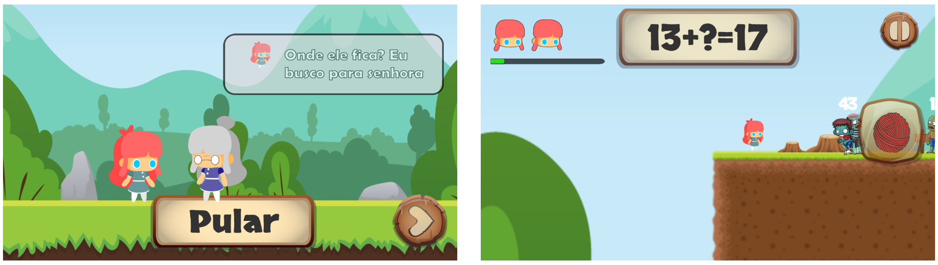
Context learning is related to the ability of models to learn from input data, classified as (i) zero-shot learning – we provide no examples in the input; (ii) one-shot learning – the prompt contains one example; and (iii) few-shot learning – we provide one or more examples in the input. Studies indicate that few-shot learning is more appropriate, showing that more examples in the prompt enhance the results [Brown et al. 2020]. For this reason, we chose to use few-shot learning in this research.

As for the model execution platform, we use integration between Ollama and AnythingLLM. Ollama is responsible for storing, managing, and making open-source models available, acting as a local execution environment accessible via API. AnythingLLM, on the other hand, allows interaction with models in an organized manner, offering support for multiple chats and conversation history, and sending files for RAG (Retrieval-Augmented Generation) tasks, among others. In this way, it acts as an interface and management layer. We chose both technologies due to their practicality and ease of use in testing models.

3.2. Game selection

The data modeling and filling of the data template, thus composing the GLA capture structure, will be based on an educational game. Therefore, this step defines which educational game and data we will send to the model. In this aspect, four selection criteria were considered: (i) the game has a scientific publication available in the literature; (ii) the existence of a gameplay demonstration video on YouTube; (iii) it includes simple mechanics, facilitating the generation of structures and the resulting analyses; and (iv) the authors' familiarity with the game. Based on these criteria, we selected the game "Trico Numérico" (Numerical Knitting).

"Numeric Knitting" focuses on helping learn the four basic mathematical operations. The story involves the young Saara, the heiress of an old family specializing in knitting. Her journey involves defeating monsters to reach the other side of the country and recover the Selmer plant, the only remaining ingredient to create an antidote that will cure her grandmother of a rare disease. The gameplay involves throwing knitting balls to defeat enemies and advance in the level. Each enemy holds a number, and the player must throw the knitting at the number that corresponds to the missing operator ("?") in the mathematical expression that appears on the screen. Figure 2 shows two moments of the game.



(a) Game storytelling.

(b) Gameplay screen.

Figure 2. Records of the game "Numerical Knitting".

Regarding the game data that make up the prompt, we chose to include: (i) game content and theme – which the game seeks to aid in learning; (ii) educational objective – what the game aims to exercise; (iii) story – detailed description of the narrative; (iv) gameplay – how the player interacts with the game; and (v) learning mechanics – how learning designers incorporated the educational content into the level design to stimulate the player's learning. Although other information about the game is also relevant, we tried not to compose a long prompt to avoid the hallucination of LLMs, which may respond imprecisely and incoherently [IBM 2023]. In addition, models tend to ignore information in the middle of the input when the prompt is long [Liu et al. 2024].

3.3. Prompt design and prompt application

This step refers to constructing the prompt to send to the models to generate the capture structures in the GLBoard template. In Honda et al. [2024], a similar study focused on developing structures by closed LLMs (ChatGPT, Gemini, and DeepSeek) under three types of context-based learning: zero, one, and few-shot learning. Given this, and considering the similarity, we used the same prompt base previously developed in the few-shot learning condition. It is worth noting that, in the previous research, we refined the prompt via the meta-prompting technique – guiding an LLM to analyze a prompt, provide feedback, and adjust it, aiming to minimize manual efforts [Ye et al. 2023] – to ensure more accurate and adequate results.

The organization of the prompt is avaliable in Figure 3, which presents five main components: (i) objective – explains the context and objective of the activity to the model;

(ii) GLBoard template – presents and details the GLA data template of the GLBoard model; (iii) steps – highlights the steps that the model must perform, emphasizing some aspects; (iv) instruction to wait – reinforces to the model that it waits for the game information to be sent (necessary, since in previous tests the models started reasoning before receiving the data); (v) examples – two capture structures filled in the GLBoard JSON template, referring to two different games; and (v) game information – sent in a second prompt to avoid a long prompt. The complete prompts are available in the following link[2].

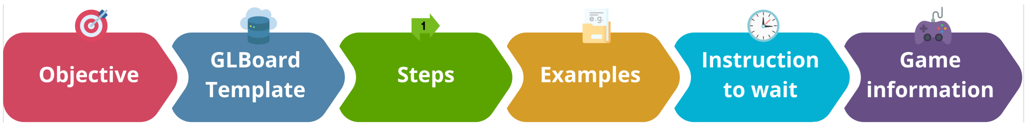


Figure 3. Prompt structure.

We applied the prompt individually using AnythingLLM and choosing one model at a time. After the first response from the models, we sent the second prompt with information about the game, Numerical Knitting. After the generation time, we recorded the responses of each model in a document for later analysis, and it is available at the link[3]. We run the models on a machine with an I7 11370H processor, 16 GB of RAM, and an NVIDIA GeForce GTX 1650 video card. The response generation time varied significantly between the models: 47 seconds (phi-4-mini-3B), 2 minutes and 43 seconds (Gemma 2-9B), 2 minutes and 31 seconds (llama3.1-instruct-8B), and 4 minutes and 47 seconds (Qwen2.5-14B). Although the models are relatively small in number of parameters, they require a lot of processing power, influencing the generation time.

3.4. Evaluation questionnaire design

Regarding the evaluation of GLA capture structures, we need to consider two points: (i) there is no instrument in the literature that validates GLA structures or data models, and (ii) the data template we used is that of GLBoard, which has its particularities. Thus, in this research, we used the ”Player Level Up!”, a form to evaluate GLA structures in the GLBoard template we designed in a previous study [Honda et al. 2025a]. It consists of seven questions, distributed in three main dimensions: (i) **Coherence**: assesses whether the names of the variables are appropriate and whether there is a clear relationship between them and the mechanics of the game, in addition to verifying whether the data types are compatible with what we intend to capture; (ii) **Redundancy and completeness**: examines whether there is overlapping information (redundant variables) and whether there is a lack of essential data that should be present in the structure; and (iii) **Evidence of evolution**: verifies whether the GLA structure allows identifying the player's progress throughout the experience, considering temporal and performance aspects.

Using Google Forms, we inserted the structures generated by the open-source models, quantitative questions (Likert-5), referring to the “Player Level Up!” instrument, and qualitative questions about each structure's negative and positive points – available at the link[4]

[2] https://drive.google.com/file/d/14Ffw9mEIsTrL9vw2QN2F3NSxd6-86Axa/view?usp=sharing
[3] https://drive.google.com/file/d/1Y85rXil2tEdCcyjeY248rH7UZwUhd5oc/view?usp=sharing
[4] https://drive.google.com/file/d/1vK-EzrNYkNySgx5-LktWkmbZ7006AiZt/view?usp=sharing

3.5. Data collection and analysis

The data we collected came from questionnaire evaluations of GLA structures. For this purpose, we invited seven experts in the field, selected for convenience due to ease of access and availability. All completed a consent form, agreeing to make their data available, anonymously, for scientific research purposes. They then completed the questionnaire and evaluated the four GLA structures, generated by the models “Gemma 2 (9B)”, “Qwen 2.5 (14B)”, “LLaMA 3.1 Instruct (8B)”, and “Phi-4 Mini (3.8B)”. We stored these data in a spreadsheet for later analysis. The profile of these experts was: (i) academic status – 71% had graduated from computer science courses and were pursuing postgraduate degrees (Master’s and Doctorate) in computer science. 29% were in advanced years (6th and 8th years) in their undergraduate degree in computer science; (ii) experience with GLA – 28.6% reported having less than 1 year of experience. This percentage of participants also reported having 1 to 2 years and 3 to 4 years of experience. Only one (14.3%) reported having more than 4 years in the field; and (iii) application of GLA in games – 3 experts (57%) indicated having already applied GLA in one game, while 29% in no match and 14% in two games.

We performed analyses based on the collected data: (i) visual analyses – boxplot with the sum of the experts’ evaluations for the models and line graph with the sum of the questions for each model; (ii) comparative analyses – with a table analyzing the models’ responses regarding the GLA activities. In addition, the models’ responses were cross-referenced with their performances in the benchmarks to assist in the analysis of the models, and (iii) content analysis [Bardin 2015] for the qualitative questions related to the positive and negative aspects of the GLA structures. We first created categories based on a preliminary reading of the responses. Then, we organized the responses into these categories and finally analyzed them together, aiming to identify general patterns and interpret the main points in each category.

4. Results and discussion

The results of this research are related to the experts’ evaluation of the GLA capture structures generated by open-source LLMs and the analysis of the responses generated by these models. Regarding the experts’ evaluations, we constructed the boxplot in Figure 4, presenting the sum of the scores assigned to the questions. The X-axis contains the names of the models, and the Y-axis corresponds to the values of the sums, ranging from 7 (all experts assigned a score of 1 to all questions) to 35 (maximum score assigned by the experts to the questions).

The graph analysis shows that the “phi-4-mini” model obtained average evaluations, which may be related to the fact that it has the smallest number of parameters (3B). On the other hand, the “Qwen 2.5 (14B)” received favorable ratings, which may also be related to the number of parameters – being four times greater than the “phi-4-mini”. “Gemma 2” presented similar results to the “Qwen”, reaching the highest maximum values, the largest median, and the smallest data dispersion. This point demonstrates a good capacity of the reasoning model for generating GLA structures since it has almost half the number of parameters as the “Qwen” and performed slightly better. As for the “LLAma 3.1 (8B)” obtained unsatisfactory results: the smallest median and the most significant data dispersion. Although “phi-4-mini” has lower minimum scores, its median is higher

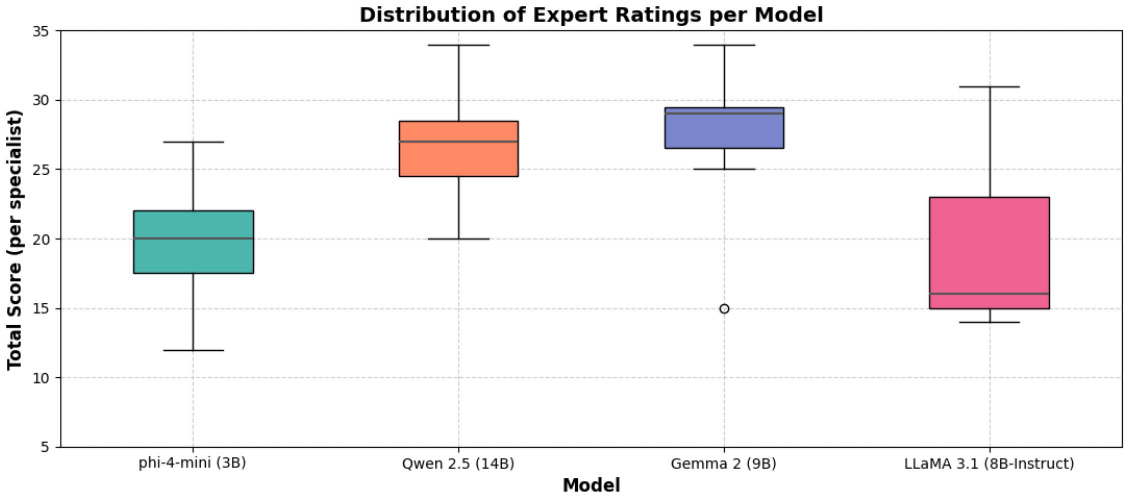


Figure 4. Boxplot graph with the sum of the experts' scores.

than that of "LLAma 3.1 (8B)", also a curious factor since the model is almost three times smaller in terms of parameters. Figure 5 illustrates the performance of the models through the sum of the scores evaluated, ranging from 7 to 35.

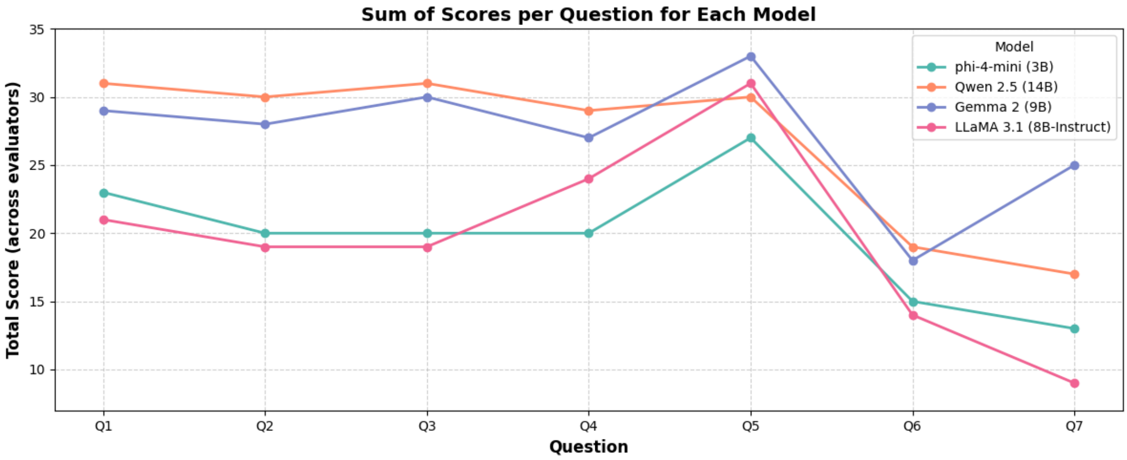


Figure 5. Line graph showing the models' scores for each question.

Regarding Q1 (variable names consistent with the data), Q2 (variables related to mechanics), Q3 (appropriate data types) and Q4 (variables belonging to the path_player), the GLA specialists evaluated the models similarly: "Qwen" with the highest scores, slightly higher than "Gemma", which also performed well, indicating consistency of the models in proposing GLA variables. "phi-4" remained superior to LLaMA in the first three questions, but both with average scores. In Q4, "LLaMA" obtained a more adequate evaluation, but still lower than "Qwen" and "Gemma". In Q5 (duplicate variables), all models presented similar scores, indicating they could propose unique variables. On the other hand, in Q6 (variables belonging to the path_player), the models presented unsatisfactory scores, indicating the difficulty of LLMs in analyzing educational games and proposing variables corresponding to the player's path to identify evidence of learning. This point supports the information that data modeling for GLA is a complex activity.

In Q7 (temporal analysis), "Qwen" and "Gemma" presented reasonable/average scores, being the only models to propose variables related to time in play.

Regarding GLA activities, the responses from each model generated the capture structures in the GLBoard template and included other aspects of the area. To analyze the quality of these responses, we made an analysis based on four criteria: (i) understanding of the game – the model interpreted the game well, citing one or more of its elements; (ii) data modeling – the model was able to list the data that the game would collect and its justifications; (iii) JSON structure – the model adequately filled the GLBoard data template; and (iv) path_player – the variables are consistent with the player's path through the levels. We assigned a value to each of these topics (excellent, good, partial, insufficient, or inadequate) (Table 4).

Model	Game Understanding	JSON Structure	Path_player	Data Modeling
phi-4-mini	Partial: mentions the content, educational goal, and gameplay, but the GLA structure showed inconsistency in some mechanic-related information	Partial: follows the model but misplaces variables and mislabels names and types	Inadequate: proposes external (aggregated) variables to the path	Partial: lists names and types, but without justifications.
qwen2.5-14B	Partial: shows understanding based on variable justification but presents no game summary	Good: coherent with the GLBoard template	Good: does not propose aggregated variables in the path and includes comments, but omits timestamps and some relevant variables	Excellent: includes variable names, data types, and detailed justification
gemma2-9B	Partial: mentions educational goal and content but does not detail gameplay or learning mechanics	Partial: misplaces variables, though still relevant to the context	Insufficient: although the variables belong to the path and include comments, they do not accurately reflect game mechanics	Excellent: the model created a data modeling table with variable names, data types, and justification
llama3.1-8B	Insufficient: superficially mentions content and educational objective	Insufficient: the model proposed few GLA variables, mostly misplaced	Inadequate: the model proposed path variables from the minimum path educational game we used as an example in the prompt, unrelated to Tricô Numérico	Inadequate: the model only listed topics, with no related proposals

Table 3. Evaluation of Models Based on Game Learning Analytics Criteria.

In response to the RQ, the models perform differently (Table 4). In **game understanding**, no model described the game clearly, only mentioning some aspects. Regarding the **JSON structure** of GLBoard, most models proposed variables in incorrect places. Some of them, although relevant, change the data template – which isn't possible. The highlight was "Qwen", which presented a structure consistent with the template. Regarding **path_player**, only "Qwen" proposed variables that were adequate and consistent with the player's path. However, it did not include other relevant variables, corroborating the experts' positive evaluations of the structure generated by the model. The others, for the most part, proposed aggregated variables that do not belong to the path. Regarding

data modeling, the "Qwen" and "Gemma" models include the names of variables, data types, and justifications for collection, respectively, in topics and tables. In addition, we compared the performance of the models in the benchmarks we used as selection criteria with the quality of the models' responses in the GLA activities, which we detailed below.

phi-4mini: we classified most of the responses as partial. The model presented the lowest performance in the MuSr benchmark (6.45), related to the ability to handle multi-step reasoning, directly reflected in its median evaluations. Furthermore, despite its reasonable score in the BBH benchmark (38.74), which focuses on natural language reasoning, this was not reflected in its responses. On the other hand, the model demonstrated a good performance in the IFEval benchmark (73.78), related to adherence to instructions, reflected in a good sequence in executing the instructions requested in the prompt. Generally, the model is suitable for well-defined tasks but presents unsatisfactory performance when it requires greater reasoning.

Qwen2.5-14B: presented the best overall evaluation, especially in data modeling. Its understanding of the game and adherence to prompt instructions were partial, consistent with its low performance in IFEval (36.94) – the weakest of all models. Its high scores in the BBH (45.08) and MuSR (15.91) benchmarks explain its positive evaluations in the reasoning activities for modeling data and composing structures. The model's analysis is the opposite of phi-4mini; although it presents difficulties in detailed instructions, it compensates with a good reasoning capacity.

llama3.1-8B: although the model did not perform satisfactorily in IFVal (49.22), it was able to follow instructions adequately. On the other hand, its responses in the GLA activities were unsatisfactory, mainly because it proposed a path_player related to another educational game. However, its low scores in the BBH (29.38) and MuSr (8.61) benchmarks justify its performance.. The performance in IFVal suggests that the model may have difficulty following instructions. However, this was not the case, and the analysis of its responses suggests that researchers can use it in contexts that require less reasoning.

gemma2-9B: presented the highest score among the models in IFEval (74.36), consistent with its good performance in following instructions. However, models with lower scores presented more appropriate instruction sequences. In the GLA activities, its performance was below expectations, mainly due to its high scores in the BBH (42.14) and MuSr (9.74) benchmarks. The model analysis shows that the reasoning was adequate to model the data, with difficulty positioning variables and proposing data that were out of tune with the game mechanics.

The models could perform GLA activities, but with several limitations. The highlight is the "Qwen" model, which obtained the most satisfactory results. "Gemma" also stands out, receiving the best evaluation from experts in the GLA structure it generated, indicating its potential, mainly because it only has 9B parameters (almost half the amount of "Qwen"). However, we notice that the models had difficulties filling the data capture structure, mainly regarding variables belonging to the path_player, which corroborates that data modeling is a complex activity. Although some models performed this activity well, they presented inconsistencies in "fitting" the data variables into the GLBoard template.

5. Conclusions

GLA enables the collection, analysis, and visualization of serious game data. However, modeling GLA data is a complex and challenging activity. Researchers have been applying LLMs to this context to minimize this difficulty, but we have found limitations of closed models. This study presents the following research question: “How do open-source LLMs perform in GLA activities, especially in filling data templates?”. To achieve this objective, we conducted an empirical study.

We followed a sequence of steps to achieve the objective of the study, involving the following steps: (i) model selection, in which we selected models “phi-4mini”, “Qwen2.5-14B”, “gemma2-9B”, and “llama3.1-8B”. The prompt technique we defined was the basic prompt and learning in context, as few-shot learning. The platform we chose to use these models was AnythingLLM; (ii) game selection: we selected the educational game “Numerical Knitting” as the basis for the GLA structures. The data we defined to send to the prompt were: content/theme, educational objective, story, gameplay, and learning mechanics; (iii) prompt design: we constructed the prompt, organized into objective, GLBoard template, steps, examples, instruction to wait and game information; (iv) prompt application: we applied the prompt to the models, whose responses we recorded in a spreadsheet; (v) construction of an evaluation questionnaire: we used the “Player Level Up!” model, creating a Google Form with its questions; (vi) data collection: seven GLA experts filled out the form, whose evaluations we stored in a spreadsheet; and (vii) data analysis: we used graphs, comparative tables, and visual analysis.

Responding to the QP, the results reveal that the models perform differently but can generate capture structures in the GLBoard template. However, they have difficulty building the capture structures, even though some have correctly modeled the data. This challenge demonstrates that modeling these data to express evidence of player learning is a complex activity. “Qwen” stood out in the other GLA activities, mainly in the generation of the structure, but GLA specialists best-evaluated “Gemma” about path_player. Most of the model’s results agree with their scores in the benchmarks, which were also an important finding in the research, being directly related to GLA activities.

Among the limitations of this research, we can mention: (i) the subjective perception of the authors, who were responsible for evaluating the responses of the GLA activities of the models, which may have presented some inconsistency in the analysis; (ii) the elaborated prompt that, although previously constructed with meta-prompting, may not yet be complete and completely adequate; (iii) the computational limitations, resulting in the choice of models with few parameters (3B to 15B); (iv) the selected benchmarks, considering only IFEval, BHH and MuSR. Including other benchmarks could have resulted in more accurate models; and (v) focus on the basic prompt technique due to its ease of use. Other techniques, such as chain-of-thought prompting [Wei et al. 2022], can enhance the models’ responses.

Future work includes (i) using a server with more robust technical configurations to allow testing with models with larger parameters, (ii) finding more benchmarks related to GLA activities, (iii) identifying more models to perform tests, (iv) including tests with RAG (Retrieval-augmented generation) to investigate the models’ capabilities more deeply; (v) performing a performance comparison with closed models; and (vi) performing tests with structures from other games.

6. Acknowledgement

The authors express their gratitude to the State University of Amazonas (UEA) for the institutional support and to PROPESP-UEA for the financial support. They are also grateful to their colleagues from the ThinkTEd Lab for the contributions and discussions that enriched this work.

References

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., and Fernández-Manjón, B. (2022). Game learning analytics:: Blending visual and data mining techniques to improve serious games and to better understand player learning. *Journal of Learning Analytics*, 9(3):32–49.

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., and Manjón, B. F. (2021). Data science meets standardized game learning analytics. In *2021 IEEE Global Engineering Education Conference (EDUCON)*, pages 1546–1552. IEEE.

Alonso-Fernández, C., Cano, A. R., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., and Fernández-Manjón, B. (2019). Lessons learned applying learning analytics to assess serious games. *Computers in Human Behavior*, 99:301–309.

Banihashem, S. K., Dehghanzadeh, H., Clark, D., Noroozi, O., and Biemans, H. J. (2024). Learning analytics for online game-based learning: A systematic literature review. *Behaviour & Information Technology*, 43(12):2689–2716.

Bardin, L. (2015). Análise de conteúdo (la reto & a. pinheiro, tradução)(6ª edição). *Lisboa, Portugal: Edições*, 70.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Din, S. U., Baig, M. Z., and Khan, M. K. (2023). Serious games: An updated systematic literature review. *arXiv preprint arXiv:2306.03098*.

Freire, M., Serrano-Laguna, Á., Manero, B., Martínez-Ortiz, I., Moreno-Ger, P., and Fernández-Manjón, B. (2016). Game learning analytics: Learning analytics for serious games. In *Learning, design, and technology*, pages 1–29. Springer Nature Switzerland AG.

Hauge, J. B., Berta, R., Fiucci, G., Manjón, B. F., Padrón-Nápoles, C., Westra, W., and Nadolski, R. (2014). Implications of learning analytics for serious game design. In *2014 IEEE 14th international conference on advanced learning technologies*, pages 230–232. IEEE.

Honda, F., Pessoa, M., Pires, F., and Harada, E. (2025a). Chatgpt, gemini or deepseek? an empirical study in game learning analytics. In *Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames)*. SBC.

Honda, F., Pires, F., Pessoa, M., and Oliveira, E. H. (2024). Building a specialist agent to assist in the implementation of game learning analytics techniques. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 2856–2865. SBC.

Honda, F., Pires, F., Pessoa, M., and Oliveira, E. H. T. (2025b). Challenges in educational game data modeling from the perspective of computing students: an empirical study. In *Workshop sobre Educação em Computação (WEI)*. SBC.

IBM (2023). Ai hallucinations. https://www.ibm.com/topics/ai-hallucinations. Accessed: 2025-06-11.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Kitto, K., Whitmer, J., Silvers, A., and Webb, M. (2020). Creating data for learning analytics ecosystems.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Lo, L. S. (2023). The art and science of prompt engineering: a new literacy in the information age. *Internet Reference Services Quarterly*, 27(4):203–210.

Macena, J., Honda, F., Melo, D., Pires, F., Oliveira, E. H., Fernandes, D., and Pessoa, M. (2024). Desafios na implementação de técnicas de gla em um jogo educacional de algoritmos: um estudo de caso. In *Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames)*, pages 814–825. SBC.

Melo, D., Melo, R., Bernardo, J. R. S., Pessoa, M., Rodríguez, L. C., and Pires, F. (2020). Uma estratégia de game learning analytics para avaliar level design em um jogo educacional. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 622–631. SBC.

Misiejuk, K., López-Pernas, S., Kaliisa, R., and Saqr, M. (2025). Mapping the landscape of generative artificial intelligence in learning analytics: A systematic literature review. *Journal of Learning Analytics*, pages 1–20.

Mitsea, E., Drigas, A., and Skianis, C. (2025). A systematic review of serious games in the era of artificial intelligence, immersive technologies, the metaverse, and neurotechnologies: Transformation through meta-skills training. *Electronics*, 14(4):649.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Shin, J., Tang, C., Mohati, T., Nayebi, M., Wang, S., and Hemmati, H. (2023). Prompt engineering or fine tuning: An empirical assessment of large language models in automated software engineering tasks. *arXiv preprint arXiv:2310.10508*.

Silva, D., Melo, R., Pires, F., and Pessoa, M. (2021). Avaliacão de objetos digitais de aprendizagem: como os licenciados em computação analisam jogos educacionais? *Revista Novas Tecnologias na Educação*, 19(2):111–121.

Silva, D., Pires, F., Melo, R., and Pessoa, M. (2022). Glboard: um sistema para auxiliar na captura e análise de dados em jogos educacionais. In *Anais Estendidos do XXI Simpósio Brasileiro de Jogos e Entretenimento Digital*, pages 959–968. SBC.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., et al. (2022). Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Van Eck, R. (2006). Digital game-based learning: It's not just the digital natives who are restless. *EDUCAUSE review*, 41(2):16.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ye, Q., Axmed, M., Pryzant, R., and Khani, F. (2023). Prompt engineering a prompt engineer. *arXiv preprint arXiv:2311.05661*.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. (2023). Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.