How do LLMs analyze and interpret data from educational games? A study with GLA experts

Manuela Bastos[1], Fabrizio Honda[1], Márcia Lima[1], Marcela Pessoa[1], Fernanda Pires[1]

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

{mcpb.snf21,msllima,msppessoa,fpires}@uea.edu.br

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

Abstract. *Game Learning Analytics (GLA) holds promise for extracting insights into player learning, especially with visual analytics techniques such as dashboards. However, data analysis and interpretation are not always straightforward. An emerging alternative is the use of Large Language Models (LLMs). In this work, we investigated ChatGPT, Gemini, DeepSeek, and GLA Specialist in analyzing data from an educational game collected by the GLBoard model. We conducted an empirical study in which two GLA experts evaluated the results. In its two tested versions, Gemini performed best, excelling in profile analysis, generating pedagogical insights, and providing detailed assessments.*

1. Introduction

Using games as learning objects to develop cognitive skills and support curricular content has intensified, as they can promote student motivation and engagement [Genesio et al. 2024, Plass et al. 2015]. Despite the benefits, the assessment of learning in educational games still shows gaps: (i) the lack of adequate instruments leads to informal (ad-hoc) assessments; (ii) the predominance of self-assessment questionnaires, which, although useful, are insufficient to identify learning; and (iii) the complexity of games which hinders the verification of educational objectives [Petri and von Wangenheim 2017, Silva et al. 2021, Pires et al. 2018].

One approach that can help in this context is Game Learning Analytics (GLA), which deals with collecting, analyzing, and interpreting data from serious games [Freire et al. 2016]. This data corresponds to records of player interactions with game elements (logs/traces), collected non-intrusively so as not to impair gameplay. Thus, through GLA, we can track players' progress through levels, collecting data such as decision-making, enemies defeated, platforms skipped, and more. After data collection, GLA models typically include Visual Analytics (VA) techniques, such as dashboards, allowing learning designers to perform analyses and gain insights from learning evidence [Few 2006, Verbert et al. 2013].

However, we noted some challenges in the VA literature, such as the predominance of descriptive analyses, which limit the extraction of deeper insights [Susnjak et al. 2022], and the cognitive overload caused by the lack of narratives that facilitate data interpretation [Liu et al. 2024], among others. Furthermore, metrics organization often lacks contextualization, which can fragment the analysis

DOI: 10.5753/sbie.2025.12887

[Macfadyen and Dawson 2012]. These challenges become even more complex in the context of educational games, which involve multiple types of data and are highly complex to create.

Generative Artificial Intelligence (GAI) can be an alternative to overcome these challenges, and researchers already use it widely in education, especially with LLMs [AlAli and Wardat 2024]: models that process natural language and generate contextualized responses [Pang et al. 2025]. In VA, especially in dashboards, GAI can reduce challenges by: (i) enabling natural language interaction through chatbots; (ii) supporting the identification of patterns and anomalies; and (iii) automating reports with explanatory narratives [Agarwal and Sonbhadra 2025, Zhao et al. 2024]. These capabilities have also motivated applications in GLA, such as creating the "GLA Specialist": an agent that assists in data modeling and filling out capture templates [Honda et al. 2024]. However, the agent focuses on pre-GLA rather than analysis or visualization. To date, no studies have combined LLMs and GLA with a focus on VA.

Therefore, this work presents the following research question (RQ): "How do the LLMs ChatGPT, Gemini, and DeepSeek perform in analyzing and visualizing GLA data from log analysis?". The main proposal is to investigate how these models can identify patterns and generate interpretations from analyzing educational game logs, contributing to Visual Analytics in GLA.

2. Foundations and Related Work

GLA, or Learning Analytics for Serious Games, results from the union of two fields: (i) Game Analytics (GA), focused on the collection, processing, analysis, and interpretation of interaction data in digital games [El-Nasr et al. 2016], generally to optimize the user experience and financial results (sales and microtransactions); and (ii) Learning Analytics (LA), focused on analyzing the behavior and learning paths of students in educational environments [Larusson and White 2014]. Therefore, GLA refers to the collection, analysis, and visualization of data from serious games [Freire et al. 2016], helping to identify evidence of learning and generating contributions to stakeholders (students, educators, and developers) by tracking the player's progress through the levels.

GLBoard is one of the models that enables the application of GLA techniques, focusing on collecting and analyzing data from educational games [Silva et al. 2022]. The authors designed their data template based on common game data, ensuring flexibility and application in different contexts, structured in JSON format. It consists of four classes: (i) PlayerData, with player profile data; (ii) GameData, with gameplay information; (iii) Phase, with data for each level; and (iv) Section, which groups sessions and includes the $path\_player$ field, responsible for recording the player's path, actions, decisions, states, and events. Furthermore, GLBoard has four modules: (i) Unity package, for integration with games developed in this engine; (ii) API, which manages the system and sends standardized data to the database; (iii) database, which stores interactions; and (iv) dashboard, which performs analysis and presents interpretable visualizations.

Whether with GLBoard or other GLA models, using Visual Analytics, such as dashboards, helps support learning designers in interpreting data. However, in more complex analyses with these resources, we note challenges in the literature: (i) a predominance of descriptive approaches, without advancing to predictive or prescriptive dimensions [Susnjak et al. 2022]; (ii) a risk of cognitive overload, especially among users with low visual literacy, due to the lack of narratives to guide interpretation [Liu et al. 2024]; (iii) difficulty in organizing and contextualizing metrics, which often appear fragmented [Macfadyen and Dawson 2012]; and (iv) dependence on the user's perspective, as cognitive biases and prior experiences can lead to divergent conclusions [Alhadad 2016, Wall et al. 2018].

Generative AI can overcome these obstacles, especially with LLMs, such as ChatGPT, which are gaining traction in everyday life and supporting content creation and decision-making tasks [Yao et al. 2024]. Recent studies show that LLMs can strengthen the field of Visual Analytics by enabling natural language queries, supporting curation and visualization generation, producing narratives that contextualize information, and adaptively incorporating domain knowledge. These models can act as "co-analysts", assisting in data analysis and interpretation and mitigating the challenges mentioned above [Hutchinson et al. 2024, Agarwal and Sonbhadra 2025].

Furthermore, researchers have also been applying LLMs to the field of GLA, as in the case of the "GLA Specialist": an agent in ChatGPT to support learning designers in modeling and customizing data capture templates, enabling data collection in games and identifying evidence of learning [Honda et al. 2024]. Although GLA experts positively evaluated it, it emphasizes pre-GLA, not data analysis or visualization. To date, we have not identified any work combining GLA and LLMs/Generative AI focusing on this stage – the related studies found below address, at most, two of these fields.

The work of Davalos et al. [2025] investigates how LLMs can support teachers by transforming multimodal reading data (such as eye-tracking, learning outcomes, and logs) into clear pedagogical reports. The study involved 82 5th-grade students in three assessments, applying unsupervised clustering techniques, with K-Means standing out, and two LLM agents – one to generate reports and the other to evaluate them. Five teachers and an LLM evaluator analyzed the reports, receiving positive reviews for clarity and usefulness, especially in the clusters and outliers sections. The research contributes to the field of VA by showing that LLMs can replace traditional dashboards with interpretive reports, bringing complex data closer to teaching practice without losing human supervision.

Alonso-Fernández et al. [2021] present T-Mon (Trace Monitor), a VA platform for serious games that automates the analysis of data collected through the xAPI-SG (Experience API for Serious Games) standard. The tool processes traces in Jupyter Notebooks and generates dashboards with seven tabs that display metrics such as progress, score, completion time, and interactions. The authors tested T-Mon with data from previous experiments, and the tool reduced technical barriers for teachers and researchers, allowing them to monitor student performance clearly and quickly. Its contribution lies in offering an accessible and standardized entry point for GLA, bringing educators closer to analyzing educational games through automated visualizations.

Zhao et al. [2024] present LEVA (LLM-Enhanced Visual Analytics), a framework that uses LLMs to support users in VA systems. It operates in three stages: during onboarding, it generates interactive tutorials; during exploration, it recommends insights based on data and system status; and during summarization, it organizes the history into automatic reports. The authors implemented LEVA in dashboards, including Tableau, and validated its effectiveness in two use cases and a study with 20 participants. The results show that assisted users achieved greater accuracy and satisfaction, and completed tasks faster than the control group. The main contribution is demonstrating that LLMs can act as mixed-initiative partners, helping users learn, explore, and document complex analyses more easily and efficiently.

Table 1 shows that this study innovates by integrating LLMs, Serious Games, Learning Analytics, and Visual Analytics. The main contributions are: (i) the systematic comparison of multiple LLMs (ChatGPT, Gemini, and DeepSeek), unlike studies that use a single tool; (ii) the focus on post-processing and qualitative analysis of logs, going beyond the rigidity of dashboards like T-Mon; (iii) the evaluation by GLA experts, based on technical and analytical criteria, in contrast to studies focused on usability; (iv) the use of meta-prompting to refine instructions and ensure consistency; and (v) the evaluation of the "GLA Specialist" in the context of data analysis and visualization, even though this is not its original focus. Thus, the study is a practical reference for using LLMs in analyzing educational game data.

Table 1. Comparison between related works and this research.

Work	LLMs	Serious Games	Learning Analytics	Visual Analytics
Davalos et al. [2025]	X		X	X
Alonso et al. [2021]		X	X	X
Zhao et al. [2024]	X			X
This work	X	X	X	X

3. Methods and Study Description

This research aims to investigate the performance of LLMs in the analysis and visualization of GLA data. To this end, we designed a methodology with six steps, illustrated in Figure 1 and described below.

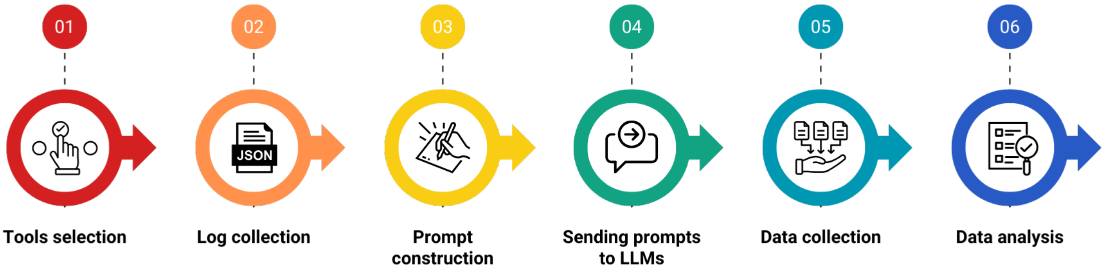

Figure 1. Research methodology.

3.1. Tool selection

This step consists of selecting the study tools, such as the GLA activity we will investigate, the educational game from which we will collect player interactions (logs), and the LLMs that will analyze the logs.

Regarding GLA, researchers generally follow these steps: data modeling, filling variables in a capture template, data collection, and visualization and log analysis. In this research, we focused on data analysis and visualization, since few studies have explored these activities at the intersection of GLA and LLMs. One example is the "GLA Specialist", which focuses not on data post-processing but on modeling and filling the template (pre-GLA).

We chose the game "Robô Euroi" (Euroi Robot) [Melo et al. 2018], which aims to aid in the learning of basic mathematical operations (addition and subtraction) and the development of Computational Thinking (CT). The game's story revolves around Euroi, a robot who must visit planets and recover the parts of a scientist's ship destroyed by an electromagnetic wave. His main objective is to repair the spaceship, avoid dangers, and defeat the enemies on these planets. The robot has limited energy, which decreases as it jumps through the platforms. When the robot collects energy, the amount increases. Therefore, the player must correctly manage the robot and guide it strategically along the paths to avoid running out of energy, considering consumption and recovery calculations. We selected this game based on the following criteria: (i) it has a scientific publication available and free in the literature; (ii) it includes a demonstration video on YouTube[1]; and (iii) the game had to be integrated with the GLBoard model and already include an implemented data template.

We selected the following LLMs: (i) ChatGPT[2], recognized for generating well-structured texts in natural dialogue, combined with the intuitive interface that consolidated it globally [Bang et al. 2023]; (ii) Gemini[3], from Google, with an emphasis on complex reasoning tasks such as programming, logic, and mathematical problem-solving [Imran and Almusharraf 2024]; and (iii) DeepSeek[4], focused on programming and open source, trained with large volumes of data [Guo et al. 2024]. We mainly considered easy access and recognition of the models as selection criteria. Furthermore, we selected two versions for each model (Figure 2): a free basic model and another focused on advanced logic and reasoning, generally available with paid plans. We also included the "GLA Specialist" in the study, which is not a model but an agent developed in ChatGPT. This inclusion is relevant because it is the only one identified in the literature that integrates GLA and LLMs [Honda et al. 2024]. We did not choose fully open-source models, since they require greater technical knowledge and could compromise time and reproducibility. Therefore, we prioritized those available on chatbot platforms, which facilitate the application of prompt engineering.

3.2. Log collection

This step involved collecting the data required for analysis, based on logs generated during player interactions with the game "Robô Euroi". As mentioned previously, the game already integrated the GLBoard architecture, and researchers had used it with different target audiences: (i) children outside the intended age range; (ii) children of the appropriate age; and (iii) adults. Although approximately 70 people participated, the game's author provided us with a single file containing logs from only five users. Because each

[1] https://www.youtube.com/watch?v=hQcWFo1EnXY
[2] https://chatgpt.com/
[3] https://gemini.google.com/u/7/app
[4] https://chat.deepseek.com/

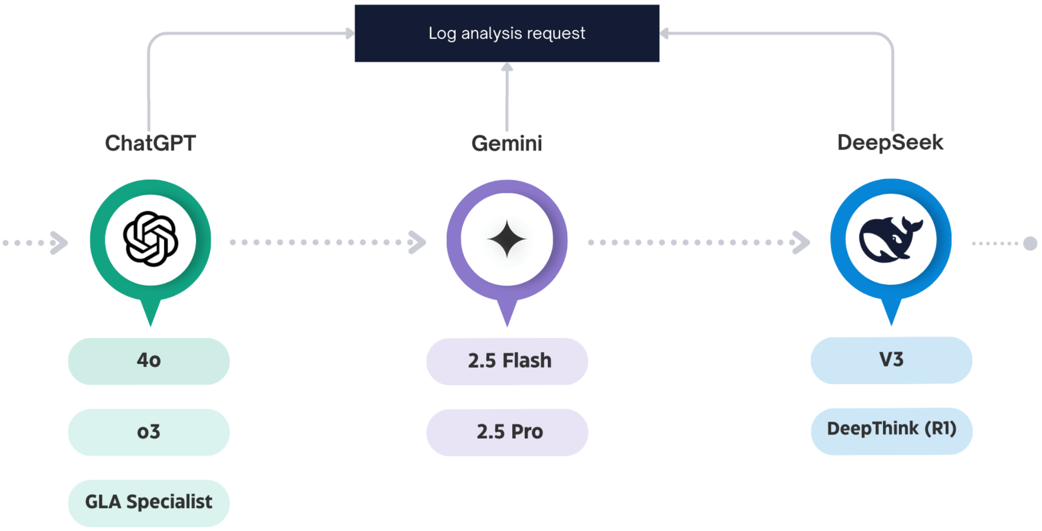

Figure 2. Models chosen for the study.

player completed multiple levels, this file contained approximately 12,755 lines, which revealed a significant volume and granularity. Such volume could have compromised analysis by LLMs due to context window limitations, computational cost, and response time. Therefore, we restricted the analysis to this dataset so that LLMs could analyze and interpret it effectively.

The logs include different GLA variables, recorded in each game level, including $bestPath$, representing the ideal path to completion (e.g., Fixed A → Plat-3 B → Fixed C), and a set of sessions (attempts). Each session stores information such as the start and end dates (e.g., 04/23/2024 19:08:01 to 04/23/2024 19:08:15); the $finalized\_challenges$, which record the challenges faced with attributes such as name (e.g., Jump: Fixed A – Plat-3 B), type (e.g., Mechanics), and status (e.g., success); the $path\_player$, which describes the platforms actually traversed (e.g., Fixed A → Fixed A → Plat-3 B); the performance, which expresses the number of stars obtained (e.g., 3); and the conclusion, which indicates the outcome of the session (e.g., Victory). We emphasize that the game's creator did not collect names or sensitive personal information, but only IDs and some basic demographic data, such as gender and birthday. Participants provided these data after completing a consent form, in which they authorized the use of their data exclusively for research purposes, which was in line with data protection guidelines.

3.3. Prompt construction

For the selected LLMs to assist in log analysis and visualization, we needed to design appropriate prompts (instructions) for this purpose, which we performed in this step. This process is prompt engineering, which consists of developing clear and contextual instructions with a well-defined format and controlled verbosity to obtain effective results from the model [Lo 2023]. We organized prompt construction into two phases: context-based and refinement.

During the context database construction stage, we developed a layered prompt to facilitate and organize the information. We divided the prompt into three parts: (i) an introduction to the field of GLA, presenting its theoretical basis and general objectives; (ii) a contextualization of the GLBoard platform, explaining its architecture and data model; and (iii) a request for analysis, specifying that it should be comprehensive and interpretive, with an emphasis on records in the $path\_player$ field and the identification of learning cues, difficulties, strategic behaviors, and interaction patterns.

The second step involved refining and optimizing the prompt. To ensure greater accuracy and clarity of the instructions, we used the “GLA Specialist” [Honda et al. 2024], since it could contribute more efficiently to the prompt’s design as an agent specialized in GLA. In this process, we applied the meta-prompting technique, which consists of developing prompts to guide the creation or interpretation of other prompts, establishing more consistent guidelines for the model [Ye et al. 2023]. The agent then corrected ambiguities, improved technical terminology, and restructured the information flow, resulting in the final prompt, available at the link[5].

3.4. Sending prompts to LLMs

In this step, we applied the refined prompt to the selected models to observe how each responded to the proposed instructions. To achieve this, we inserted the final prompt into new, independent conversations with the selected LLM versions, ensuring uniform execution and comparison conditions. During this process, we identified an initial inconsistency in ChatGPT. Instead of waiting for the log file, the model immediately generated an analysis based on dummy data, which compromised the usefulness of the response. To overcome this limitation, we added explicit instructions at the end of the prompt, requiring the model to start the analysis only after receiving the file.

We then applied the adjusted prompt to all models and sent the log file. It’s important to note that DeepSeek reported analyzing only 52% of the data in the file. The other models didn’t report this limitation; however, we assumed they also didn’t process the entire content, since the logs contained over 12,000 lines. This behavior highlights a limitation of closed models, which have limited analysis capabilities when subjected to large volumes of data.

3.5. Data collection and analysis

In this step, we collected data corresponding to the responses the LLMs provided after we submitted the prompt. To organize the results, we stored all responses in an online document[6], totaling seven distinct outputs. We identified each response with the model name and its respective version, for example, “Gemini 2.5 PRO.” For this research, we analyzed only the first responses each model generated, aiming to investigate their initial interpretations of the presented prompt. Therefore, we chose not to conduct additional interactions in the conversations, since our focus was to analyze how the LLMs understand and immediately respond to requests related to GLA data.

Two GLA experts analyzed the data collected, namely the responses the LLMs generated from the prompt, using ChatGPT as a supporting tool. The first expert holds a

[5] https://drive.google.com/file/d/1mNqhLGIHLyldGn4NNS8eoeZY66b21rpq/view?usp=sharing

[6] https://drive.google.com/file/d/1lPtidh7aVcARqEJX8U-5iM6M7kPsr3gR/view?usp=sharing

Bachelor's degree in Computer Science Education from the State University of Amazonas (UEA) and is currently pursuing a Master's degree in Computer Science at the Federal University of Amazonas (UFAM), with eight years of experience in educational games and three years of direct experience in GLA. The second expert holds a Bachelor's degree in Information Systems from UEA and has two years of experience in educational games and GLA. Both experts are authors of this work.

Thus, we used two complementary approaches to interpret and evaluate the model responses. The first was a content analysis [Bardin 2015], in which experts initially read the responses to identify central ideas. From these, they defined thematic categories and grouped the responses according to these categories, which allowed us to recognize recurring patterns and highlight relevant points within each group. The second approach consisted of constructing a comparative table, indicating which GLA criteria each model met. Additionally, we created a summary table with the main pros and cons observed for each model.

4. Results and Discussions

We presented the results of this research on two fronts: (i) the individual analysis of the responses the LLMs generated, conducted through content analysis; and (ii) the preparation of comparative tables that summarize the performance of the models based on GLA data analysis and visualization criteria.

Regarding individual analyses, **ChatGPT-4o** stood out for generating qualitative insights into the learning process, such as strategic adaptation after mistakes and overcoming challenges through trial and error. The limitation lies in its focus on a single player and unclear visual data, such as the path graph without identifying the levels. Even so, the approach remains aligned with the GLA by using the $path\_player$ variable as an indicator of difficulties and opportunities for pedagogical and game design intervention.

ChatGPT-o3 stood out for its comprehensive analysis and pedagogical approach, focusing on aggregated behavioral patterns. Using metrics such as success rate and number of attempts, it highlighted the escalation of difficulty in the final stages and the contrast between high performance in mechanical challenges and low performance in logic tasks. Its main merit was translating this data into pedagogical recommendations, such as contextual explanations and reflective activities, associating longer trajectories with evidence of learning.

Gemini 2.5 Flash presented the most sophisticated analysis aligned with GLA principles, standing out for integrating individual and collective perspectives. The model developed detailed profiles of each player, inferring cognitive processes such as resilience and adaptation based on paths taken, types of errors, and strategic evolution. At the same time, it identified collective patterns, such as difficulties in specific mechanics and critical phases, transforming these findings into practical recommendations. Suggestions ranged from adjustments to game design and tutorials to pedagogical strategies, such as the use of data to promote collaborative learning. By linking performance with concrete actions for educators and developers, the response proved to be a comprehensive and applied analysis.

Gemini 2.5 Pro presented a structured and comprehensive analysis, grouping players into behavioral profiles such as "Explorers" and "Strategists" based on performance indicators. Using the variable $path\_player$, it highlighted critical phases of the learning process, highlighting understanding of feedback, overcoming difficulties, and abandonment due to frustration, demonstrating how interaction data reveals cognitive processes. Its main merit was translating these findings into pedagogical recommendations, proposing personalized interventions for each profile and suggesting concrete actions for educators and developers. Thus, the response demonstrated sophistication in using data not only for diagnosis but also to personalize the learning process and improve game design.

DeepSeek presented a clear and straightforward analysis, profiling three players and identifying difficulty trends, such as the sudden increase in complexity in Stage 10 and the recurring challenges in logic and platforming mechanics. While it provided practical recommendations for balancing and tutorials, the analysis lacked depth, as it did not explore the $path\_player$ variable in detail or offer data visualizations. Thus, it functioned more as a diagnostic tool for game design than as an in-depth pedagogical tool, although it demonstrated a good general understanding of the data and GLA principles.

DeepSeek DeepThink (R1) presented a quantitative and comparative analysis, identifying player profiles and critical difficulty points based on performance metrics. However, the absence of the $path\_player$ variable prevented the exploration of cognitive processes and trial-and-error strategies, making the approach more descriptive than interpretative. Although it provided practical recommendations, it functioned primarily as a statistical diagnostic, which is less useful for pedagogical applications that require an in-depth understanding of player behavior.

The **"GLA Specialist"** presented a deep qualitative analysis focused on a single player, demonstrating how to extract insights even from limited data. By interpreting the $path\_player$ field, it mapped the user's learning curve, from initial exploration to strategic execution, revealing cognitive processes such as adaptation, perseverance, and strategy refinement. Its strength was translating this analysis into specific pedagogical recommendations, such as contextual feedback and reflective prompts. Thus, it stood out as a model for individual analysis in GLA, demonstrating how a player's trajectory can support detailed diagnoses and didactic interventions. Despite its depth, its limitation lies in its focus on a single player, failing to explore collective patterns or multiple profiles, which limits its scope of application.

Table 2 presents the evaluation of the models based on criteria associated with the GLA activities, which focus on data analysis and visualization. Each row in the table corresponds to a specific criterion that the experts defined with the support of an LLM. Columns R1 to R7 represent the evaluated responses: (i) R1 refers to ChatGPT-4o; (ii) R2 to ChatGPT o3; (iii) R3 to Gemini 2.5 Flash; (iv) R4 to Gemini 2.5 Pro; (v) R5 to DeepSeek; (vi) R6 to DeepSeek DeepThink (R1)[7]; and (vii) R7 to the response prepared by the "GLA Specialist". For this analysis, we applied a four-level scale: 4 (Completely Meets), 3 (Meets), 2 (Partially Meets), and 1 (Does Not Meet).

Regarding the criteria associated with data analysis and visualization in the context of GLA, it is observed that the strengths were in "Trial and error identification" and "Focus on individual analysis (micro)" (22 points each), which demonstrate the ability

[7] The suffix "R1" in the DeepSeek name refers to the model version, not to column R1 in the table.

Table 2. Evaluation of LLMs in GLA analysis and visualization criteria.

Criterion	**R1**	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**Total**
$Path\_player$ field analysis	3	2	4	4	2	3	3	**21**
Trial and error identification	3	3	4	4	2	3	3	**22**
Level analysis (difficulty, success, etc.)	1	2	3	3	2	2	3	**16**
Player profiles or clustering	1	1	4	4	1	1	3	**15**
Pedagogical insights generation	2	3	4	4	2	1	3	**19**
Game design recommenations	1	1	4	4	3	3	1	**17**
Didactic suggestions for teachers	1	3	4	4	1	1	3	**17**
Multiple players coverage	1	3	4	4	2	2	1	**17**
Graph generation	3	4	1	1	1	1	1	**12**
Table generation	2	4	1	4	1	3	3	**18**
Focus on individual analysis (micro)	3	2	4	4	3	3	3	**22**
Focus on aggregated patterns (macro)	1	3	4	4	2	2	1	**17**
Objective alignment	2	4	3	3	1	2	2	**17**
Total	**22**	**35**	**47**	**42**	**28**	**26**	**35**	**-**

of LLMs to recognize basic patterns of progress. Intermediate results appeared in criteria such as “Level analysis” (16 points), “Objective alignment” (17 points), and “Pedagogical insights generation” (19 points), indicating specific advances, but still limited in interpretative clarity. The weaknesses, on the other hand, were concentrated in “Player profiles or clustering” (15 points), “Multiple players coverage” (17 points), and “Graph generation” (12 points), revealing difficulties in expanding the analysis to collective levels and in translating data into visual representations. Some of these weaknesses may be associated with partial processing of logs, which may not be fully analyzed by the models, as demonstrated by DeepSeek. These findings show that, although LLMs have the potential to enrich data analysis in educational games, they still face barriers to advancing in pedagogical, collective, and visual dimensions, requiring new integration strategies with GLA practices.

The model performance analysis showed that response R3 (Gemini 2.5 Flash) obtained the highest overall score, with 47 points, demonstrating greater consistency in the GLA-related criteria. This model excelled in level analysis, player profiling, micro and macro focus, and objective alignment, in addition to generating good pedagogical insights. Next, response R4 (Gemini 2.5 Pro) obtained 42 points, also demonstrating robust performance. Responses R1 (ChatGPT-4o) and R7 (“GLA Specialist”) tied with 35 points, indicating intermediate performance, with strengths in trial and error identification, micro focus, and pedagogical recommendations. Response R5 (DeepSeek), with 28 points, performed slightly better than the lowest-scoring models, but still below the top-rated ones. Finally, R2 (ChatGPT-o3, 26 points) and R6 (DeepSeek DeepThink R1, 22 points) had the lowest scores, highlighting limitations such as limited visual analysis or a lack of interpretive depth. We also observed that, except for ChatGPT, the basic versions consistently outperformed the reasoning-focused variants.

The results support the RQ, showing that the evaluated LLMs have the potential to support data analysis and visualization in GLA from educational game logs, although with differences in depth and consistency. Overall, the Gemini models stood out as the most robust, while ChatGPT-4o and the "GLA Specialist" occupied an intermediate position, and ChatGPT-o3 and the DeepSeek variants presented greater limitations. We also observed that, with the exception of ChatGPT, the basic versions outperformed the reasoning-oriented ones, suggesting that greater architectural sophistication did not, in this case, result in better performance. Table 3 complements this analysis by summarizing the main strengths and weaknesses of each model, highlighting contributions such as profiling and pedagogical recommendations, as well as weaknesses related to the lack of visualizations and interpretative depth.

Table 3. Comparative table with the pros and cons of each LLM.

Model	Strength	Weakness
R1	Identifying recurring paths and trial and error	Graphics without context (e.g., levels) and limited scope (1 player)
R2	Comparison between challenge types (logic vs. mechanics)	Lack of concrete visual analysis (missing graphics)
R3	Level-by-level analysis and well-defined individual profiles	Extensive text and little focus on visualization
R4	Generation of contextualized pedagogical insights	Little emphasis on graphics or visual representations
R5	Clear summary of difficulties by challenge type	No use of pathplayer and complete absence of visualizations
R6	Statistical comparison between players	Descriptive focus, without visualizations or in-depth interpretation
R7	Good micro-reading of actions and trial and error	Focus shifted to JSON structure and lack of visualizations

5. Conclusions

In this study, we investigated the potential of LLMs to support Visual Analytics practices in GLA by interpreting logs from the "Robô Euroi" game collected with GLBoard. We constructed a structured prompt and refined it through meta-prompting, applying it to versions of ChatGPT, Gemini, and DeepSeek. For each LLM, we selected two variants – one basic and one focused on reasoning – to compare differences in analytical depth. We also included the "GLA Specialist" in observing its performance outside its original focus. Two GLA specialists evaluated the responses based on analysis and visualization criteria, considering individual and collective perspectives, to identify interaction patterns, cognitive processes, and pedagogical recommendations, thereby addressing the RQ.

Our results show that, although all models have potential, their consistency varies significantly. The best-performing criteria were "Trial and error identification‘" and "Focus on individual analysis (Micro)", which demonstrated the LLMs' ability to recognize basic interaction patterns. Criteria such as "Level analysis", "Objective alignment", and "Pedagogical insights generation" achieved intermediate results, showing occasional advances but still limited interpretive clarity and depth. By contrast, "Player profiles or clustering", "Multiple player coverage", and "Graph Generation" scored the lowest, indicating difficulties in collective analysis and visual representations. In this context, Gemini performed best overall, ChatGPT ranked in the middle, and DeepSeek performed worst, especially in its reasoning variants. We also observed that, except for ChatGPT, the basic versions outperformed the reasoning-oriented ones, suggesting that greater architectural sophistication did not, in this case, result in richer analyses.

The main contributions of this study are: (i) comparison of multiple LLMs (ChatGPT, Gemini, and DeepSeek) in both basic and reasoning variants; (ii) use of meta-prompting to refine instructions and improve consistency; (iii) inclusion of the "GLA Specialist" as a support agent and comparison object; (iv) emphasis on post-processing logs to extract pedagogical evidence beyond rigid visualizations; (v) proposal of analysis and visualization criteria related to GLA, with expert evaluation; and (vi) indication to educators that LLMs can serve as guides, supporting the identification of patterns and the planning of pedagogical interventions.

As limitations of this research, we highlight: (i) use of mostly closed LLMs, without exploring open-source models; (ii) evaluation by two GLA experts only, which may have introduced interpretative biases; (iii) partial processing of logs by some models, which did not fully analyze the dataset; (iv) definition of the evaluation criteria by the authors, assisted by an LLM, which limits their external validation; and (v) construction of the prompt itself, which, even with meta-prompting, may not have represented the best strategy to explore the full potential of the models.

As future work, we plan to: (i) increase the number of experts in the evaluation process; (ii) incorporate customizable open-source models; (iii) analyze methods to handle larger volumes of data (logs) more consistently; and (iv) expand the investigation to different educational 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

Agarwal, N. S. and Sonbhadra, S. K. (2025). A review on large language models for visual analytics. *arXiv preprint arXiv:2503.15176*.

AlAli, R. and Wardat, Y. (2024). Opportunities and challenges of integrating generative artificial intelligence in education. *International Journal of Religion*, 5(7):784–793.

Alhadad, S. (2016). Attentional and cognitive processing of analytics visualisations: Can design features affect interpretations and decisions about learning and teaching? In *ASCILITE 2016*. Australasian Society for Computers in Learning in Tertiary Education (ASCILITE).

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.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

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

Davalos, E., Zhang, Y., Srivastava, N., Salas, J. A., McFadden, S., Cho, S.-J., Biswas, G., and Goodwin, A. (2025). Llms as educational analysts: Transforming multimodal data traces into actionable reading assessment reports. *arXiv preprint arXiv:2503.02099*.

El-Nasr, M. S., Drachen, A., and Canossa, A. (2016). *Game analytics*. Springer.

Few, S. (2006). *Information dashboard design: The effective visual communication of data*. O'Reilly Media, Inc.

Freire, M., Serrano-Laguna, Á., Manero Iglesias, 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: An international compendium of theory, research, practice, and policy*, pages 3475–3502. Springer.

Genesio, N. O. S., de Oliveira, A. M., Oliveira, E. W., and Valle, P. H. D. (2024). Panorama de estudos sobre jogos educacionais digitais em educação em computação. In *Workshop sobre Educação em Computação (WEI)*, pages 737–749. SBC.

Guo, D., Zhu, Q., Yang, D., Xie, Z., Dong, K., Zhang, W., Chen, G., Bi, X., Wu, Y., Li, Y., et al. (2024). Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

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.

Hutchinson, M., Jianu, R., Slingsby, A., and Madhyastha, P. (2024). Llm-assisted visual analytics: Opportunities and challenges. *arXiv preprint arXiv:2409.02691*.

Imran, M. and Almusharraf, N. (2024). Google gemini as a next generation ai educational tool: a review of emerging educational technology. *Smart Learning Environments*, 11(1):22.

Kim, M., Kim, S., Lee, S., Yoon, Y., Myung, J., Yoo, H., Lim, H., Han, J., Kim, Y., Ahn, S.-Y., et al. (2024). Llm-driven learning analytics dashboard for teachers in efl writing education. *arXiv preprint arXiv:2410.15025*.

Larusson, J. A. and White, B. (2014). Learning analytics. *From Research to Practice. Nueva York: Springer*.

Liu, Y., Pozdniakov, S., and Martinez-Maldonado, R. (2024). The effects of visualisation literacy and data storytelling dashboards on teachers' cognitive load. *Australasian Journal of Educational Technology*, 40(1):78–93.

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.

Macfadyen, L. P. and Dawson, S. (2012). Numbers are not enough. why e-learning analytics failed to inform an institutional strategic plan. *Journal of Educational Technology & Society*, 15(3):149–163.

Melo, D., de Sousa Pires, F. G., Melo, R., and Júnior, R. J. d. R. S. (2018). Robô euroi: Game de estratégia matemática para exercitar o pensamento computacional. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educaçao-SBIE)*, volume 29, page 685.

Pang, R. Y., Schroeder, H., Smith, K. S., Barocas, S., Xiao, Z., Tseng, E., and Bragg, D. (2025). Understanding the llm-ification of chi: Unpacking the impact of llms at chi through a systematic literature review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Petri, G. and von Wangenheim, C. G. (2017). How games for computing education are evaluated? a systematic literature review. *Computers & education*, 107:68–90.

Pires, F. G. d. S., Melo, R., Machado, J., Silva, M. S., Franzoia, F., and de Freitas, R. (2018). Ecologic: um jogo de estratégia para o desenvolvimento do pensamento computacional e da consciência ambiental. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 7, page 629.

Plass, J. L., Homer, B. D., and Kinzer, C. K. (2015). Foundations of game-based learning. *Educational psychologist*, 50(4):258–283.

Silva, D., Melo, R., Pires, F., and Pessoa, M. (2021). Avaliacão de objetos digitais de aprendizagem: como os licenciados em computação analisam jogos educacionais? *RENOTE*, 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 *Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames)*, pages 959–968. SBC.

Susnjak, T., Ramaswami, G. S., and Mathrani, A. (2022). Learning analytics dashboard: a tool for providing actionable insights to learners. *International Journal of Educational Technology in Higher Education*, 19(1):12.

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., and Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, 57(10):1500–1509.

Wall, E., Blaha, L. M., Paul, C. L., Cook, K., and Endert, A. (2018). Four perspectives on human bias in visual analytics. In *Cognitive biases in visualizations*, pages 29–42. Springer.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

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

Zhao, Y., Zhang, Y., Zhang, Y., Zhao, X., Wang, J., Shao, Z., Turkay, C., and Chen, S. (2024). Leva: Using large language models to enhance visual analytics. *IEEE transactions on visualization and computer graphics*, 31(3):1830–1847.