Prompting LLMs for Game Learning Analytics: A Case Study with Computer Science Students

Fabrizio Honda[1,2], **Fernanda Pires**[1], **Elaine Harada**[2], **Marcela Pessoa**[1]

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

[2]Postgraduate Program in Computer Science (PPGI)
Institute of Computing – Federal University of Amazonas (IComp-UFAM)

{fabrizio.honda,elaine}@icomp.ufam.edu.br,{fpires,msppessoa}@uea.edu.br

Abstract. *The mandatory inclusion of computer science in schools requires graduates capable of acting as learning designers. The development of educational games can aid this process by allowing the implementation of Game Learning Analytics (GLA) techniques. However, modeling data to list GLA variables is complex, motivating the use of Large Language Models (LLMs). This case study examines how a technique inspired by the chain-of-thought approach, combined with the inclusion of more game data in the prompt, affects the quality of GLA structures. Results indicate that the method is relevant; however, students struggle to guide the LLM. Furthermore, the excess of data in the prompt may have led the model to ignore important information.*

1. Introduction

According to Law No. 14.533/2023, the inclusion of computer science content and other digital competencies in the curricula of Elementary and High School education in Brazil was made mandatory [Brasil 2023]. In this context, the formation of qualified educators becomes essential. The Degree in Computing Education establishes that these professionals should be able to create tools, methods, and techniques to bring computing concepts into schools, acting as learning designers [Brasil 2016]. To develop these skills and competencies, students take higher education courses that involve creating learning objects, such as educational games [Oliveira et al. 2019, Miguel et al. 2025].

Through educational game design, students act as learning designers and develop multidisciplinary skills and competencies that are fundamental to their formation [Wu and Wang 2012, Honda et al. 2020]. At State University of Amazonas (UEA), the course "Educational Software Foundations (ESF)" enables students to build educational games that aim to minimize learning difficulties. This process ranges from brainstorming to development in a game engine and implementing techniques to collect player interaction data, known as Game Learning Analytics (GLA) [Freire et al. 2016].

GLA enable the collection, analysis, and interpretation of player data during gameplay to identify evidence of learning. The analysis of this evidence provides insights to educators, game designers, and students, which are essential for improving the game [Banihashem et al. 2024]. However, studies show that students, when acting as learning designers, face difficulties in modeling GLA data that effectively reflect learning indicators [Honda et al. 2025c]. The data modeling guides the definition and organization of

GLA variables in a capture structure (data template) that the developers will implement to collect player interactions. Therefore, it is a fundamental and complex activity.

An alternative to minimize this complexity is the use of Large Language Models: generative AI models capable of performing numerous tasks like humans, returning responses in natural language [Kasneci et al. 2023]. Research has revealed the potential of these models for GLA activities, both for data modeling and generation of capture structures, as well as for the analysis and interpretation of data logs [Filho et al. 2025, Bastos et al. 2025]. When it comes to LLMs, the amount of information we provide as input and the prompt techniques (sets of instructions) can result in more coherent and contextualized responses, such as the multi-step reasoning technique known as "chain-of-thought prompting" [Wei et al. 2022].

In this regard, this study poses the following research question (RQ): "How do a prompt technique inspired by chain-of-thought and the number of game elements in the input influence the quality of data capture structures generated by an LLM in the GLBoard template?" We conducted a case study to address this question, involving undergraduate students in Computing Education at UEA. The objective is to evaluate a strategy that may assist them in data modeling for GLA, a relevant competence in learning designer formation. Our main contributions lie in the intersection of GLA and LLMs, a topic that remains underexplored in the literature. The study encompasses data modeling for educational games, the generation of GLA structures both manually and via LLM, their expert evaluation, a comparison between basic prompting and a questioning-oriented technique inspired by chain-of-thought, and other aspects.

2. Foundations and related work

GLBoard is one of the models that enables the implementation of GLA techniques, assisting in the capture and analysis of data from educational games [Silva et al. 2022]. The goal of this model is to be generic and replicable to any educational game. To this end, it provides a data template (capture structure) with four main classes: (i) $PlayerData$ – player information (name, gender, date of birth, and ID); (ii) $GameData$ – gameplay data, which contain one or more phases; (iii) $Phase$ – the game levels, containing name, status, and a set of sessions; and (iv) $Session$ – corresponds to each attempt in a phase, including start/end date, completion, performance, and the player's path ($path\_player$).

The $path\_player$ variable represents the player's path within the phases: decision-making, time records, interaction with interfaces, among others. The developer must list the GLA variables in $path\_player$ that correspond to the specificities of their game, performing the data modeling. However, modeling data for educational games is not a trivial task, which has led to studies that use LLMs to help reduce this complexity and guide the learning designer [Honda et al. 2025c]. These models are capable of generating natural language texts based on the input (prompt), and the prompt technique can enhance their responses. Chain-of-thought prompting is one of the most relevant techniques in the context of LLMs, involving multiple steps that enable the model to "reason" [Wei et al. 2022]. In this way, we assume that incorporating intermediate strategies for LLMs, such as a chain-of-thought approach, can lead to more relevant responses.

Honda et al. [2025b] conducted an empirical study to evaluate how LLMs (ChatGPT, Gemini, and DeepSeek) respond to theoretical GLA questions and generate capture

structures (GLBoard standard). The authors tested the models in both standard and complex reasoning versions, using basic prompting under zero-, one-, and few-shot learning conditions. Each LLM answered eight questions and generated three structures, which GLA experts then evaluated. Results suggest that Gemini performed best on theoretical questions, while ChatGPT stood out in structure generation under few-shot conditions, followed by DeepSeek. The study reveals the potential of LLMs in the GLA field, though with caveats, especially regarding the omission of variables in the generated structures.

Liu et al. [2025] proposed a hybrid approach that integrates GPT-4o and XGBoost (a machine learning algorithm) to detect student struggle behaviors in the game "Wake: Tales from the Aqualab". They collected Game logs, converted them into text replays, and analyzed them using GPT-4 with different prompting strategies. The study examined over 40,000 sessions from 19,186 players between January and May 2024. Results indicate that GPT excelled at simple tasks and uncommon narratives, while XGBoost performed better with quantitative patterns. The hybrid approach outperformed the individual models, contributing to the GLA field by combining LLMs with machine learning.

In the study by Henkel et al. [2025], the authors investigated the use of LLMs to interpret open-ended math responses. They proposed AMMORE, a dataset containing over 53,000 answers from K–12 students. The study evaluated GPT-3.5 Turbo and GPT-4 in the task of scoring complex responses that a rule-based classifier was unable to label. To guide the model in evaluating the answers, the author adopted chain-of-thought prompting and zero- and few-shot learning conditions. They use a subset of the most challenging responses (1%) and compare the performance of the LLMs with human grading. Results show that the chain-of-thought technique achieved higher accuracy on this subset, demonstrating the potential of LLMs in the field of Learning Analytics, especially when using interactive prompting techniques.

Although the study by Honda et al. [2025] investigated different LLMs for GLA data modeling and compared zero-, one-, and few-shot learning, it focused solely on basic prompting. In Liu et al. [2025], although the authors utilized LLMs in a context related to GLA (without explicitly using this term), the study did not address data modeling or the completion of capture templates. The research by Henkel et al. [2025] focused on the use of LLMs within Learning Analytics, examining chain-of-thought prompting, but not in the context of educational games. Therefore, the innovative aspects of this study involve the use of LLMs to support GLA-related activities, such as data modeling and the completion of capture structures. The GLA structures are created manually by computing students and artificially via prompting with ChatGPT. We also compare basic prompting with a questioning-oriented technique inspired by the chain-of-thought approach, gather students' perceptions of the study, and invite GLA experts to evaluate the structures.

3. Methods and development

In this paper, we investigate how (i) a prompt technique inspired by chain-of-thought and (ii) a large volume of game elements in the prompt influence the quality of GLA capture structures (GLBoard template). In this case study, the researcher does not take on an active role, aiming instead to analyze a contemporary phenomenon in its real-world context through multiple data collection methods [Wohlin 2021].

3.1. Goal, context and participants

Learning designers can create GLA capture structures manually (without Generative AI) or by using LLMs. Research shows that these strategies are not mutually exclusive, but rather complementary [Honda et al. 2025a]. This study aims to investigate how the quality of these structures varies depending on the use of a questioning technique inspired by chain-of-thought prompting and the inclusion of a large number of game elements in the prompt. In this context, we analyze three approaches (Figure 1): A1 – manual modeling by computing students acting as learning designers; A2 – modeling via LLM using basic prompting, conducted by a GLA expert (author of this paper); and A3 – modeling via LLM using the questioning technique, carried out by the same students. The expert has more than six years of experience with educational games and more than one year with GLA, holding a degree in Computing Education, and currently pursuing a master's degree in Computer Science at the Federal University of Amazonas (UFAM).

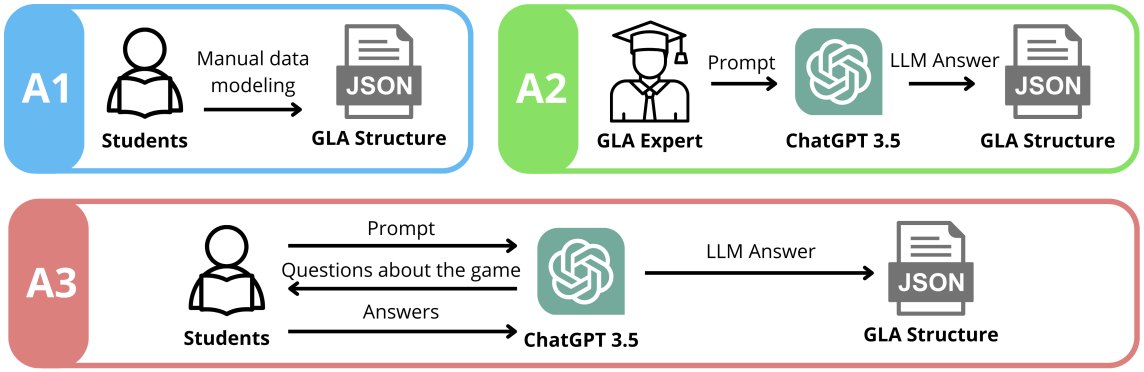


Figure 1. Approaches we investigated in the study.

The context of this research relates to the ESF course, part of the Computing Education degree program at UEA. In this course, students need to create educational games to minimize the learning problems they identified. We selected the participants based on convenience sampling, specifically students currently at the data modeling stage, where they are defining GLA variables and filling in capture structures (the contemporary phenomenon under investigation). Nine computing students participated of the study, of whom 90% enrolled in the Computing Education degree program and 10% were pursuing a Bachelor's in Information Systems[1]. The students were in their 6th semester (40%) or beyond the 8th semester (60%), and none had prior experience with data modeling.

3.2. Procedures

We conducted the procedures of this study in seven stages (Figura 2) as described below.

LLM selection: Involves selecting which model we will use to generate the GLA structures. We chose ChatGPT, as its application in education is well established [Ansari et al. 2024]. We selected version 3.5 because, at the time of the study, version 4 was not freely available. Regarding the prompt technique, we chose: (i) basic prompting – due to its simplicity in composing and submitting an input to the LLM; and (ii) an original technique inspired by chain-of-thought prompting, given its positive results in Wei et

[1]Although not part of the Education program, we didn't exclude this student from the sample due to prior experience in game development and for taking the course as an elective, acting as a learning designer.

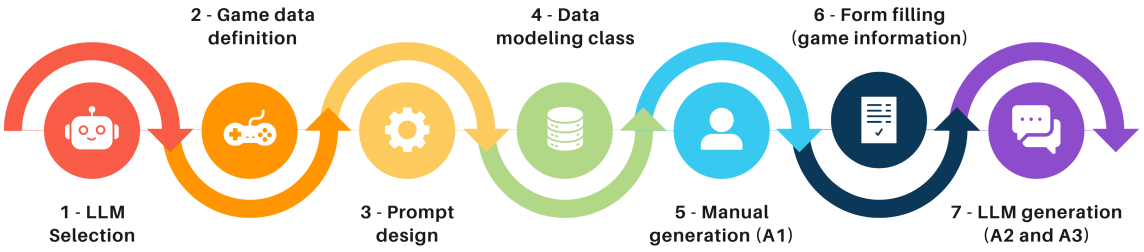


Figure 2. Steps we conducted in this study.

al. [2022]. Our technique includes an intermediate stage in which the model asks the user several questions about the educational game before generating the GLA structures. This stage encourages the learning designer to reflect on their game and learning objectives, while also providing more context to the model. We assigned the use of this technique to the students to also assess whether they had adequate knowledge of their games and how they formulated GLA prompts for the LLM. As for In-Context Learning (ICL) – model's learning based on the input – we opted for zero-shot learning, aiming to analyze solely the model's response based on its internal knowledge, without examples in the prompt.

Game data definition: Refers to which game elements we will include in the prompt sent to the LLM, so that it can return the corresponding GLA structures. In internal analysis, we identified that five elements in the prompt were sufficient to generate satisfactory structures (target audience, educational content, story, gameplay, and learning mechanics). Therefore, by increasing the number of these elements, we can provide more context to the LLM and obtain more appropriate responses. Based on this, we used the Educational Game Design Document (EGDD) [Pires 2021][2] and selected 26 elements, of which nine are mandatory for prompt composition: game name, theme, content, educational objective, target audience, story, hero's journey, game world, and gameplay. All elements and their descriptions are available[3].

Prompt design: Involves elaborating on the prompt to submit to the model by the GLA expert, which contains information about the educational games. Based on tests with ChatGPT, the resulting prompt includes four main components: (i) context – briefly introducing GLA and stating that the we will provide the GLBoard template; (ii) GLBoard data template – in JSON format, including the GLA variables, data types, and an explanation of each attribute. We chose GLBoard because the authors had prior experience with the model; (iii) objectives – detailing the step-by-step task, especially for generating GLA variables related to the $path\_player$; and (iv) conclusion – instructing the model to wait for the game data and specifying that the output structure must be in JSON format. This final component was necessary because, in some tests, the LLM did not return the answer in the expected format and included variables unrelated to the $path\_player$. The full prompt is available via the following link[4].

Data modeling class: This step refers to a class conducted by GLA experts (authors of this study) for the students. This step was necessary since the students had no prior experience with data modeling, and we contextualized the process as an activity

[2]A document that details game design elements and learning aspects.
[3]Link to the EGDD elements: https://doi.org/10.6084/m9.figshare.30789749
[4]Prompt link: https://doi.org/10.6084/m9.figshare.30789908

within the course. We conducted the class covering the following topics: (i) the concept and importance of data modeling; (ii) practical examples; (iii) alerts and tips for modeling data; (iv) definition, architecture, and data template of the GLBoard model; and (v) conclusion, description of the manual data modeling activity (A1), and references.

Manual generation of structures: The students began the manual data modeling (A1) – shown in blue in Figure 1. To do so, we required they to: (i) analyze the educational games they were designing in the course (primarily about Mathematics and Portuguese Language, and one of Computing), understanding their learning objectives; (ii) define which aspects of the player's path would be essential to identify evidence of learning; (iii) fill out an online spreadsheet provided by the experts, specifying: the data number, the name of the GLA variable, the data type, and the justification for its collection; and (iv) insert the GLA variables into the GLBoard data template, focusing on the $path\_player$ to identify signs of learning in the game phases. As a result, eight GLA capture structures were created by the students (four out of the ten participants worked in pairs on the same game). After completing the activity, they filled out a custom form (Google Forms) to submit their GLA structures and describe their perceptions of the data modeling process and the development of the structures[5].

Form filling: This stage refers to a second form completed by the students, in which they provided information about their educational games. The form fields correspond to the elements described in Step 2. This process was necessary so that the expert could generate the capture structures using ChatGPT (in a subsequent step). To assist the students with their own prompting, we also share the spreadsheet resulting from the form responses with them.

LLM structure generation: Corresponds to the use of ChatGPT to generate the GLA structures, both by the expert using basic prompting (A2) – in green in Figure 1, and by the students using the custom questioning technique (A3) – in red in Figure 1. The expert used the spreadsheet filled out by the students to insert the elements of each game into the prompt developed in Step 3. Based on this, in separate conversations with ChatGPT, the expert submitted the prompts one by one and saved the responses (GLA structures). Meanwhile, in a second class session, the students were guided by the experts to design prompts that would enable ChatGPT to receive the educational game data and ask questions about it, thereby improving its understanding of the games. The students then crafted their prompts, submitted them individually to ChatGPT 3.5 on their own devices, and responded to the questions asked by the model. Once they completed the activity, they filled out a third form[6], in which they submitted links to their interactions with ChatGPT and shared their perceptions about the process (prompt design, model responses, and comparison with manual modeling).

3.3. Data collection and analysis

The procedures we carried out produced a total of 24 GLA capture structures, with eight from each Approach. To verify whether these structures were appropriate, they needed to be evaluated by experts in the field using an assessment instrument. For this purpose, we invited three GLA experts, selected by convenience and based on their experience in

[5]Available at: https://doi.org/10.6084/m9.figshare.30789935.
[6]Available at: https://doi.org/10.6084/m9.figshare.30789944.

the area: 100% had more than six years of experience with educational games and over one year with GLA, and all held degrees in Computing Education – two were Master's students in Computer Science, and one was a PhD candidate.

Regarding the evaluation, considering the specificity of this research, which lies at the intersection of GLA and LLMs, and involves the GLBoard template, we chose to use the "Player Level Up!" instrument [Honda et al. 2025b]. This instrument assesses GLBoard GLA structures across three dimensions, comprising seven questions, as shown in Table 1. The dimensions operationalize principles of data modeling, especially regarding the definition and organization of variables related to player interactions. In this context, we created three forms (Google Forms), one for each approach. The sections of the forms correspond to the GLA structures, with each structure containing seven Likert-scale (1 to 5) questions from the instrument, plus one open-ended question for optional comments. Each expert evaluated all 24 structures.

Table 1. "Player Level Up!" evaluation instrument.

Dimension	No.	Question
Coherence	1	Variable names are consistent with the data, meaning it is clear what each variable refers to
	2	The variables defined in the structure are directly and clearly related to the game mechanics
	3	The data types assigned to each variable appear to be appropriate for what they are intended to capture
	4	The defined variables belong to player interactions (path_player)
Redundancy and completeness	5	There are no variables capturing overlapping/duplicated information
	6	No variables that could be important to capture were identified as missing from the structure
Evidence of Evolution	7	The structure allows for temporal analysis of the path, such as time spent on each stage or task, or timestamps

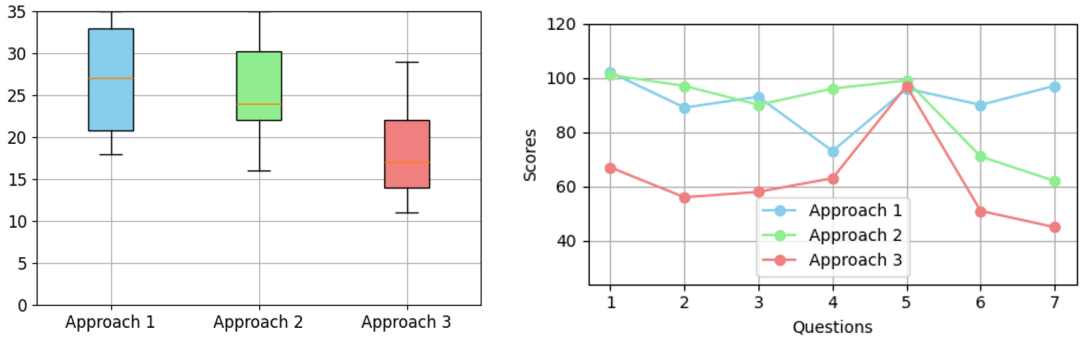
It is worth noting that, for all forms we applied, both the students and GLA experts authorized the anonymous use of their data solely for research purposes. Furthermore, this study ensures confidentiality and complies with research ethics, having been approved by the Research Ethics Committee under CAAE number 85800424.8.0000.5020, in accordance with the substantiated opinion no. 7.465.500.

A case study must involve multiple methods of data collection [Wohlin 2021], which we included: (i) empirical data – students' interactions with ChatGPT 3.5; (ii) objective measures – quantitative evaluations from students regarding both manual data modeling and the use of LLMs, as well as the prompting process for GLA, in addition to expert evaluations of the generated structures; (iii) subjective perspectives – qualitative feedback from students about the GLA capture structures generated; and (iv) documentary evidence – the GLA capture structures generated both manually and through interactions with ChatGPT. We performed data triangulation to enhance the validity of the study.

For data analysis, we used: (i) a boxplot comparing the approaches based on the experts' quantitative evaluations of the structures; (ii) a line chart displaying the evaluations of each approach by question from the "Player Level Up!" instrument; (iii) descriptive statistics, presenting the medians, standard deviations, and minimum/maximum values of the approaches; (iv) a stacked bar chart, quantifying the GLA variables proposed across the three approaches, distinguishing between atomic variables and those belonging to the $path\_player$; and (v) thematic analysis for the qualitative questions. Based on an initial review of the responses, we created thematic categories to group them. We then assigned the responses to their corresponding categories to highlight patterns and insights.

4. Results and discussion

Figure 3 presents two charts: (i) a boxplot showing the experts' assessment of the capture structures. The X-axis represents the approaches, while the Y-axis shows the total evaluation scores for the structures (with a minimum score of 7, where all questions for a structure received the lowest rating from a expert, and a maximum of 35, where all questions received a rating of 5); and (ii) a line chart, where the X-axis corresponds to the questions from the "Player Level Up!" instrument and the Y-axis represents the total evaluation scores, ranging from 24 (if all three experts gave the minimum score to all eight structures of a given approach) to 120 (maximum score from all three experts).



(a) Boxplot of the evaluated approaches.

(b) Distribution of scores per question.

Figure 3. Results of the evaluations conducted by the experts

Through the boxplot in Figure 3(a), we found that Approach 1 (A1), in which students manually created the capture structures (without Generative AI), was the most highly rated by the GLA experts. Approach 2 (A2), where a GLA expert generated the structures via ChatGPT using basic prompting, achieved satisfactory results similar to those of A1, though slightly lower. Approach 3 (A3), in which ChatGPT generated the structures based on student-designed prompts using a questioning-oriented technique, proved to be the least satisfactory.

Regarding the questions, Q1, Q2, and Q3 refer, respectively, to consistent variable names, data related to game mechanics, and appropriate data types. While Approaches 1 and 2 received satisfactory evaluations on these items, indicating consistency in the proposed GLA variables, Approach 3 did not reach adequate scores (all below the average), suggesting weaknesses in variable definition when using LLMs guided by student-designed prompts. Q4 addresses the player's progression through the game phases ($path\_player$), where experts rated Approaches 1 and 3 unsatisfactorily. This assessment highlights the challenges faced by learning designers, both in manual modeling and in guiding an LLM, in defining data aligned with their educational games. The positive evaluation of Approach 2 on this item suggests that, when properly instructed, the model is capable of generating suitable GLA variables. In Q5, which concerns the presence of duplicate variables, all approaches performed well, indicating consistency in proposing unique variables. Q6, focused on essential variables not being included, again highlights modeling difficulties. Approaches 2 and 3, involving LLMs, received unsatisfactory scores, suggesting limitations in the model's comprehensiveness. Finally, regarding Q7 (temporal analysis), experts gave acceptable scores for A1, but rated A2

and A3 as low. This point reinforces the notion that LLMs often overlook key variables related to the player's timeline or action duration during gameplay.

Table 2 presents the descriptive statistics of the evaluated Approaches, including the median, standard deviation (SD), minimum/maximum values, and range. We can observe that the evaluations of Approaches 1 and 2 were similar, particularly in terms of median, minimum/maximum values, and range, reinforcing the proximity between these two approaches. The lowest standard deviation is associated with Approach 2, indicating greater consistency among the experts in their evaluations. Overall, Approach 3 performed worse than the others and showed a consistent standard deviation. This point suggests agreement among the experts in assigning lower scores, indicating that the approach was unsatisfactory. Although the highest score in Approach 3 was 29, it refers to a structure that was highly rated and created by a student with experience in programming and games. These scores, however, were not consistent across the other structures, suggesting that this case was an outlier.

Table 2. Descriptive Statistics of Scores by Approach

Approach	Median	SD	Min	Max	Range
Approach 1	27.0	6.14	18	35	17
Approach 2	24.0	5.27	16	35	19
Approach 3	17.0	5.34	11	29	18

Regarding the GLA structures in the three approaches, we conducted an analysis, with support from ChatGPT, to identify whether the GLA variables belonged to the $path\_player$ or consisted of atomic (or aggregated) variables, which are calculated post-phase and do not represent the player's path. This analysis resulted in Figure 4, where we use the LLM solely to assist with the interpretation and organization of the results. The Y-axis contains the eight educational games created, with each bar representing the structures generated according to the approach (A1, A2, or A3). Meanwhile, the X-axis shows the number of defined variables, organized by lighter tones (atomic variables) and darker tones (variables from the $path\_player$).

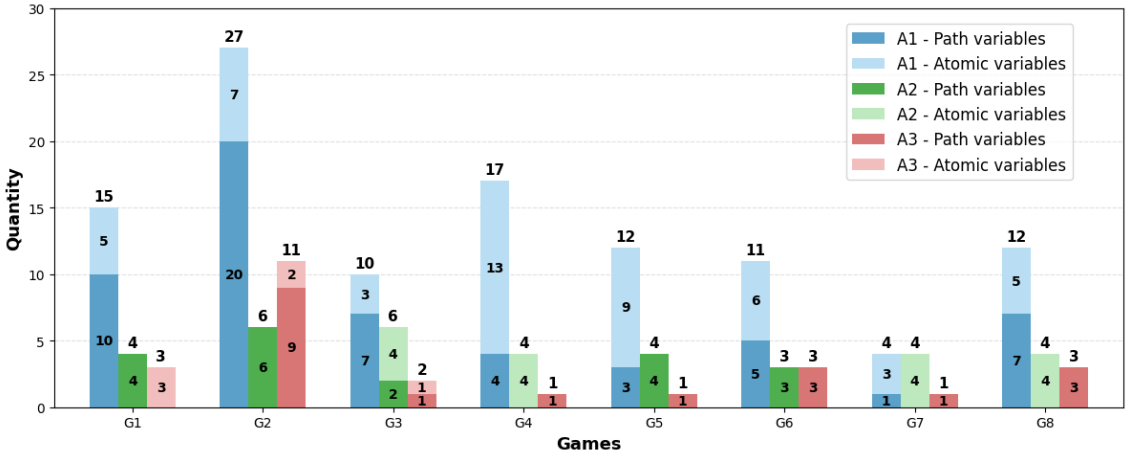


Figure 4. Stacked bar chart with the quantity of variables.

The chart provides a clearer understanding of the evaluations. In Approach 1, students manually identified a total of 108 GLA variables, of which 52.8% belonged to the $path\_player$. Except for G2, most of the capture structures in A1 included between ten and seventeen variables. Although this quantity was satisfactory, we observed that many variables were not properly organized within the data template, particularly in distinguishing them from atomic variables. This limitation also created challenges in interpreting the variables used to generate the chart. For example, in player-enemy encounters, we noted the use of variables like $enemy : int$, instead of representing the encounter as an event object, such as "enemy X encountered at position (X, Y) at HH:MM:SS". These issues highlight not only the complexity of modeling data for educational games but also the difficulty in properly structuring GLA variables within the data template. The game G2 stood out with the highest number of total variables and $path\_player$ variables. A student with experience in game development and programming was responsible for the game, as reflected in the strong evaluation of their capture structure.

In Approach 2, ChatGPT defined a total of 35 GLA variables, representing only 32% of the number proposed in Approach 1. Of these, 54.2% belonged to the $path\_player$. We observed weaknesses in three of the eight structures, where the model failed to identify any variables related to the player's path and suggested only atomic variables. On the other hand, in four cases, it proposed only $path\_player$ variables and no atomic ones. These results indicate that the model has the potential to generate contextualized responses, but also struggles with consistency and standardization. In Approach 3, only 25 variables were proposed, with a higher proportion (76%) belonging to the $path\_player$. However, 47% of these came from a single game (G2), while the other games had at most three variables each. In one case, the model generated over 29 variables, but only one was actually related to the player's path. In two other cases, the LLM proposed only a single GLA variable. We also identified some inconsistencies in the LLM's output, such as including examples of variable values instead of specifying the data types. These issues align with the experts' evaluations for A2 and A3 in Q6, which concern essential GLA variables not being considered. This point indicates persistent challenges for the model in adequately identifying and defining relevant variables.

Regarding the students' perceptions of the study (measured through qualitative and quantitative questions), we excluded one student from the sample because he did not complete the first evaluation form. 44% (n=4) found that manually modeling data was "easy", 44% considered it "average", and 12% (n=1) found it "difficult". They thought the most challenging aspects to be: defining the data and their justifications (44%), organizing the data (33%, n=3), and ensuring the data reflected the player's learning progression (23%, n=2). Regarding the organization of data in the template, 67% (n=6) found it "easy", with the most significant difficulty encountered being with the $path\_player$ variable. Although students perceived this step as easy, the structure analyses revealed it to be one of the biggest challenges, mainly due to confusion between atomic variables and those related to the player's path. Regarding the difficulty in composing the questioning-oriented prompt for ChatGPT, 45% (n=4) found it "average", 33% (n=3) found it "easy", and 22% (n=2) considered it "difficult". The main challenges listed were: structuring what ChatGPT needed to generate (45%), limitations of the model (22%), and responding to the model's questions (11%, n=1). Regarding the understanding of which data would be essential to collect in relation to the educational game and why, 45% believed

that ChatGPT helped with this process. Additionally, 78% (n=7) reported that one or two separate conversations with the model were necessary to obtain a satisfactory result.

Although experts rated low scores to Approach 3, the majority of students (78%, n=7) agreed that the questioning helped in the construction of the capture structure, especially in terms of: "rethinking the characteristics of the game" (P3), and "because of the well-directed questions" (P6), among others. When asked about the use of ChatGPT compared to manual modeling, 78% of the students considered it a good idea, mainly in terms of understanding what needs to be collected. P6's response stands out in this context: "GPT is an excellent tool for correction and improvement, not for creation". This point reinforces one of the key goals of the research: the role of Generative AI is not to replace the learning designer, but to assist them in modeling data and designing capture structures in a guided way and with supportive tools. Students' preferences regarding the structure generation technique support this point, which shows that both have their advantages: 45% (n=4) prefer ChatGPT, 33% (n=3) felt that the manual approach was more appropriate, and 22% (n=2) remained neutral.

In response to the RQ, the use of questioning and the number of game elements present in the prompt influence the quality of GLA capture structures in different ways.

Questionments : The research reveals signs of the potential of questioning for generating more adequate capture structures (78% of the students agreed, n=7). Although 78% of the students did not consider composing the prompt "difficult" and 45% (n=4) did not identify missing variables for collection, Approach 3 showed the most unsatisfactory results, especially regarding the absence of variables. This point suggests that students encounter difficulties in understanding and perceiving the quality of capture structures. Therefore, we cannot state that the questioning worsened the responses, but rather that the students' inexperience influenced the structures in designing prompts (they knew the model but had not taken a class on prompt design) and implementing GLA techniques.

Inclusion of more game data in the prompt: We observed that this factor may have hindered the model's ability to generate the structures, mainly due to the absence of GLA variables whose descriptions were present in the prompt elements. In this context, the large volume of data may have harmed the quality of the structures more than the students' inexperience itself. What supports this claim is the strong evaluation of Approach 2, in which the GLA expert created a well-organized prompt, despite ChatGPT not proposing many variables. This analysis aligns with the findings of Liu et al. [2024], who report that LLMs tend to ignore information from the prompt, especially in the middle. These characteristics underscore the importance of a well-crafted prompt or the use of a strategy that does not rely on complex prompts, such as an intelligent agent.

The study presents evidence that designing prompts for GLA is challenging, even for students who have previously used ChatGPT. On the other hand, it also suggests potential for LLMs in the field: GLA specialists have well evaluated Approach 2. This result is particularly relevant in the context of the Degree in Computing Education, where students act as learning designers in courses and need to apply knowledge from the field to model data from educational games, especially from subjects such as programming, systems modeling, and databases [Honda et al. 2025c]. However, this stage is not trivial, reinforcing the importance of this study in investigating how LLMs can support the use

of this knowledge, assist the learning designer, and improve GLA structures.

We highlight as limitations: (i) the short duration of the classes, which did not cover prompt engineering concepts for the students, potentially affecting the quality of the structures; (ii) the use of ChatGPT 3.5, which is limited compared to versions 4 and 4o, but we chose for being free; (iii) the small sample size, as not all participants completed every step or filled out the forms (we excluded these outliers); (iv) potential biases in the experts' understanding of the games when evaluating the structures; (v) the variable chart, which, although experts created it with support from ChatGPT, may not have correctly represented the belonging of GLA variables to the $path\_player$; and (vi) the simultaneous analysis of two factors (questioning and additional game data in the prompt), which we didn't isolate in separate studies and may have affected cause-and-effect inferences.

5. Conclusions

This case study examined how the quality of GLA capture structures (GLBoard standard), varies depending on the use of a questioning-oriented technique inspired by chain-of-thought prompting and the inclusion of additional game elements in the prompt. We compared three approaches: A1 – manual creation of the structures by students; A2 – artificial generation of the structure by a expert using ChatGPT with basic prompting; and A3 – artificial generation of the structure by students using ChatGPT with the questioning technique. We invited experts in the field to evaluate the structures.

Overall, none of the GLA experts evaluated the structures with scores below 10, suggesting that they were minimally adequate. A1 was the most satisfactory, demonstrating that the manual generation of structures by students was superior. A2 showed less data dispersion, with good agreement among the experts, and had median and minimum/maximum values similar to those of A1. A3 was the lowest-rated, and students had mixed opinions about their own structures: 45% (n=4) identified missing variables, while 45% disagreed. This point also reveals a challenge in students' perception, since A3 received the lowest evaluation in Q6 (missing variables). Although the results are promising, there is still room for improvement. As noted by P9, "Chat can sometimes hinder more than help, returning data that is inconsistent with the structure".

In response to the RQ, questioning can be helpful; however, the students were unable to guide the LLM effectively. This point may be related to their lack of experience with prompt engineering, despite not perceiving it as a difficulty. The inclusion of more data in the prompt may also have negatively influenced the quality of the structures, as the LLM did not propose many GLA variables. On the other hand, the research enables the identification of a strategy that may assist in data modeling for educational games, a fundamental and challenging stage for learning designers. In this way, it provides relevant evidence for the Degree in Computing Education, since prior knowledge from courses in the field contributes to more consistent modeling. As future work, we intend to: (i) include prompt engineering lessons for students; (ii) analyze the prompts they create; (iii) invite more GLA experts to evaluate the structures; and (iv) explore other LLMs.

6. Acknowlegment

We carried out this study with the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (AUXPE-CAPES-PROEX) – Finance Code 001. Ad-

ditionally, this work was partially funded by the Amazonas State Research Support Foundation – FAPEAM – through the PDPG-CAPES project. It also received support from the National Council for Scientific and Technological Development – CNPq (Process 303443/2023-5).

The authors express their gratitude to the State University of Amazonas (UEA) for the institutional support. They are also grateful to their colleagues from the ThinkTEd Lab for the contributions and discussions that enriched this work. The authors further acknowledge the support of the Nexus research center, which provided essential material resources, including physical space, computers, and related infrastructure.

Use of Generative Artificial Intelligence (GAI)

In this study, we used ChatGPT from OpenAI to generate the codes for graphs in Overleaf, aiming to help minimize time and effort in constructing these representations.

References

Ansari, A. N., Ahmad, S., and Bhutta, S. M. (2024). Mapping the global evidence around the use of chatgpt in higher education: A systematic scoping review. *Education and Information Technologies*, 29(9):11281–11321.

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.

Bastos, M., Honda, F., Lima, M., Pessoa, M., and Pires, F. (2025). How do llms analyze and interpret data from educational games? a study with gla experts. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1317–1330. SBC.

Brasil (2016). Resolução nº 5, de 16 de novembro de 2016.

Brasil (2023). Lei nº 14.533, de 11 de janeiro de 2023. Diário Oficial da União. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/lei/l14533.htm. Acessed: 2025-12-01.

Filho, D., Honda, F., Pires, F., Pessoa, M., et al. (2025). Exploring the use of open-source llms for game learning analytics: an empirical study. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1302–1316. SBC.

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.

Henkel, O., Horne-Robinson, H., Dyshel, M., Thompson, G., Abboud, R., Ch, N. A. N., Moreau-Pernet, B., and Vanacore, K. (2025). Learning to love llms for answer interpretation: Chain-of-thought prompting and the ammore dataset. *Journal of Learning Analytics*, 12(1):50–64.

Honda, F., Pessoa, M., Harada, E., and Pires, F. (2025a). Evaluation of a specialist agent in game learning analytics by learning designers: a case study. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1361–1375. SBC.

Honda, F., Pessoa, M., Pires, F., and Oliveira, E. H. (2025b). Chatgpt, gemini or deepseek? an empirical study in game learning analytics. In *Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames)*, pages 1828–1840. SBC.

Honda, F., Pires, F., Pessoa, M., and de Oliveira, E. H. T. (2020). Liçoes aprendidas em computaçao através da criaçao de um jogo educacional: entre automatos e design de aprendizagem. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1753–1762. SBC.

Honda, F., Pires, F., Pessoa, M., and Oliveira, E. H. (2025c). Challenges in educational game data modeling from the perspective of computing students: an empirical study. In *Workshop sobre Educação em Computação (WEI)*, pages 1251–1262. SBC.

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.

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, X., Wei, Z., Baker, R. S., Metcalf, S. J., Zhang, J., Barany, A., Slater, S., Swanson, L., and Gagnon, D. J. (2025). Integrating large language models and machine learning to detect struggle in educational games. In *International Conference on Artificial Intelligence in Education*, pages 398–405. Springer.

Miguel, J., Macena, J., Honda, F., Pires, F., and Pessoa, M. (2025). Robohouse: incorporating level and learning design into the playful approach to data structures. In *Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames)*, pages 1758–1769. SBC.

Oliveira, M. G., Santos, R. F., and Pereira, L. L. (2019). Jogos digitais educacionais: um mapeamento sistemático da literatura brasileira. *Revista Brasileira de Informática na Educação*, 27(1):123–145.

Pires, F. G. d. S. (2021). Thinkted lab, um caso de aprendizagem criativa em computação no nível superior.

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.

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.

Wohlin, C. (2021). Case study research in software engineering—it is a case, and it is a study, but is it a case study? *Information and Software Technology*, 133:106514.

Wu, B. and Wang, A. I. (2012). A guideline for game development-based learning: a literature review. *International Journal of Computer Games Technology*, 2012.