ChatGPT, Gemini or DeepSeek? An empirical study in Game Learning Analytics

Fabrizio Honda[1,2], **Marcela Pessoa**[1], **Fernanda Pires**[1], **Elaine H. T. Oliveira**[2]

[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,{msppessoa,fpires}@uea.edu.br

Abstract. Introduction: *Generative artificial intelligence (GenAI), through large language models (LLMs), is an emerging area that has transformed the way of solving problems in several areas of knowledge. This technology can also offer solutions to challenges still open in Game Learning Analytics (GLA). Modeling data for GLA and implementing data capture and analysis techniques in educational games to identify evidence of learning are not trivial tasks. The use of LLMs can bring benefits; however, the wide variety of models available, which work in different ways, makes choosing the most appropriate model challenging.* ***Objective:*** *In this context, this article presents a comparative analysis of how LLMs (ChatGPT, Gemini, and DeepSeek) perform activities related to GLA, including the generation of templates for data capture in the GLBoard model under zero, one, and few-shot learning conditions.* ***Methodology:*** *For this, an empirical study was conducted with steps that involved the selection of models, definition of questions, construction of prompts, and collection and analysis of data.* ***Results:*** *The results indicated that ChatGPT and DeepSeek presented more accurate responses under the few-shot learning condition, with DeepSeek standing out in both GLA activities.*
Keywords *Generative Artificial Intelligence, Large Language Models, Game Learning Analytics, Educational Games, Prompt Engineering.*

1. Introduction

Educational games have been increasingly highlighted as tools capable of promoting motivation and engagement, thus facilitating learning. However, assessing developing skills and competencies through games is still challenging. The area of Learning Analytics applied to Serious Games, also called Game Learning Analytics (GLA), emerges as an alternative to identifying the development of skills and competencies through player interaction in a stealthy manner [Freire et al. 2016, Banihashem et al. 2024]. Thus, by tracking player interactions (traces), data collection and analysis are carried out to identify their evolution during gameplay [Freire et al. 2016, Alonso-Fernandez et al. 2017].

Despite its benefits, the GLA area presents some obstacles, such as (i) the implementation of its techniques, which is complex and costly [Saveski et al. 2016, Alonso-Fernández et al. 2022]; (ii) the lack of standardization, where many games have their capture strategies, which are not replicable [Alonso-Fernandez et al. 2017]; and (iii)

studies generally do not contemplate in detail the data modeling process, which involves defining which data will be captured and the justification of why they are relevant to help identify evidence of learning. To minimize the challenges related to the implementation complexity and lack of standardization, Silva et al. [2022] proposed GLBoard. This tool includes a flexible data capture model, allowing the storage of variables standard to games (player data, phases, sections, etc.) and data specific to any educational game. Regarding the lack of details on data modeling, more recently, contributions can be obtained from the area of Generative Artificial Intelligence (GAI), which has been encouraging studies in several areas of knowledge, dealing with generative modeling and deep learning to create texts, images or other forms of media [Jovanovic e Campbell 2022, Sengar et al. 2024].

The emergence of the GAI area occurred mainly with the advancement of Large Language Models (LLMs): models trained on robust databases capable of generating text like humans and performing tasks with high accuracy [Kasneci et al. 2023]. The widespread use of these models was boosted by the emergence of ChatGPT [Brown et al. 2020], which currently has more than 14,000 related publications in Scopus (source: author). In addition, other models have also gained notoriety, such as (i) Google's Gemini, which performs well in programming activities and is the first model to outperform human experts in Massive Multitask Language Understanding (MMLU), one of the most popular methods for testing AI models in knowledge and problem-solving skills [Imran e Almusharraf 2024] and (ii) DeepSeek, in operation since January 2025, considered the "rival" of ChatGPT, whose performance is comparable, but at a relatively lower cost [BBC News 2025].

Using GAI as a possibility to solve some open problems in GLA, Honda et al. [2024] proposed "GLA Specialist", a conversational agent created in ChatGPT to assist in data modeling and implementation of GLA techniques according to the GLBoard model. However, different LLMs can bring different results, directly impacting modeling and variable selection for GLA in an educational game. Thus, this study investigates how the LLMs ChatGPT, Gemini, and DeepSeek perform in generating data capture structures in the GLBoard template by conducting an empirical study. Two versions of each LLM were used for a more in-depth analysis, and zero, one, and few-shot learning in the prompts informed the models. Therefore, this study presents a research question: "How do ChatGPT, Gemini, and DeepSeek perform in specific GLA activities?".

2. Background

GLA techniques make it possible to track players' paths, promoting an understanding of their decision during the phases and allowing the generation of insights into learning progression. In this way, we have an evidence-based approach that collects and analyzes player interaction data (trace data), providing input to teachers, developers, and students.

GLBoard [Silva et al. 2022] is a model for capturing and analyzing data from educational games. Integrated with the Unity game engine, the model can be incorporated into games by installing the package, which allows developers to access the GLBoard data template and fill it out according to their game. The data in this template is flexible for any educational game: both standard variables are common to any game, for example, player data, phases, sections, and fields such as $path\_player$, where sets of variables specific to the game in question can be defined to store the player's path. Since its

structure is customizable, filling it out is more efficient when the data is correctly modeled beforehand. However, this is not a trivial process – learning designers have difficulties. An alternative to overcome this obstacle may be using LLMs, whose applicability to various domains has been increasing.

The advancement of LLMs has impacted several sectors of society and has been studied by numerous researchers, especially ChatGPT [Liu et al. 2023b, AlBadarin et al. 2023]. The capabilities of these models have made them applicable to any context, whether in general or specific domains. In this regard, training LLMs become essential, enabling them to be customized in several domains and making them more effective and efficient [Nexocode 2023]. Fine-tuning is one of the main approaches to training LLMs, in which the model parameters are adjusted according to a specific training data set, allowing its customization to the respective domain [Nexocode 2023]. Performing fine-tuning, however, requires that the model's source code, such as LLaMA (open source), be available. However, fine-tuning is not possible in closed models like ChatGPT or Gemini since the model weights are not publicly available. In these circumstances, an alternative to train them is through prompt engineering.

Prompt engineering refers to the design of input instructions (prompts) to the LLM so that they are effective in achieving the desired result [Lo 2023]. This strategy is less costly and simpler than fine-tuning. Still, it also has limitations, such as the complexity of creating prompts and the challenge in contexts that require specific knowledge, such as medicine, computing, and academic settings [Zamfirescu-Pereira et al. 2023, Lo 2023]. Regarding prompt techniques, typologies and taxonomies have sought to systematize them [Liu et al. 2023a, Sahoo et al. 2024]; however, given the emergence of the area and the diverse contexts, it is not possible to determine which technique is most appropriate, as it depends on its particularities. In-Context Learning (ICL) can assist in elaborating more effective prompts, which concern the behavior/learning of the model from the input. It can be classified into three types: (i) zero-shot learning, where only the prompt is informed, without any examples; (ii) one-shot learning, in addition to the input, an example is informed to the model; and (iii) few-shot learning: two or more examples are included in the prompt, in addition to the description. Studies suggest that few-shot learning obtains more adequate results, whose examples assume the characteristic of "prior knowledge," enhancing the model's responses by considering it [Brown et al. 2020, Mann et al. 2020, Chan et al. 2023].

Given the possibilities of LLMs, their implications are also being studied in domains such as Game Learning Analytics (GLA) [Freire et al. 2016, Honda et al. 2024]. The first (and only, as far as research has been done) study on the intersection of these two areas was that of Honda et al. [2024], which proposes and introduces the "GLA Specialist", an intelligent agent created in ChatGPT, which assists in the pre-and post-implementation of GLA techniques. However, despite being a customizable version, it only contemplates ChatGPT. It is unknown whether using different models, such as Gemini or DeepSeek, can obtain better results. This, therefore, is the objective of this study: to investigate the contributions of LLMs to GLA, seeking to analyze the responses of the ChatGPT, Gemini, and DeepSeek models to conceptual questions of GLA and the resulting capture structures in filling the GLBoard data template, using zero, one and few-shot learning.

3. Methods

This work is an empirical study aiming to to analyze the performance of LLMs when answering Game Learning Analytics questions and performing tasks under zero, one, and few-shot learning conditions. The steps are illustrated in Figure 1 and described below.

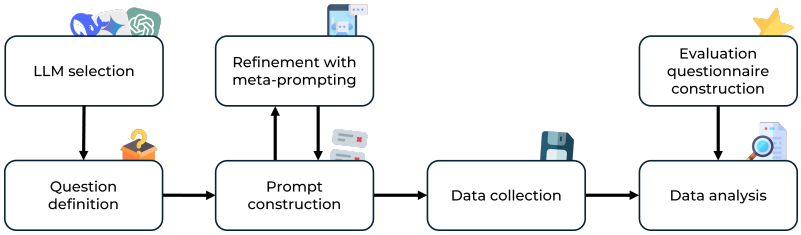


Figure 1. Study steps.

LLM selection: this step involves selecting which models would be used in the study and which ChatGPT, Gemini, and DeepSeek were listed. The justification was given because ChatGPT is the most used model globally, with more than 180 million monthly users [DemandSage 2025], being applied to several domains [Singh e Singh 2023]; Gemini performs well in tasks with long and complex chains of reasoning, standing out in programming, logic, reading, mathematical problems and code generation [Akter et al. 2023, Imran e Almusharraf 2024]. Deepseek, due to its recent rise, is commonly compared to ChatGPT in terms of performance and with a lower cost, which is estimated to have reached the mark of 5 to 6 million global users, with just over a month in activity [SEO.ai 2025, BBC News 2025]. Furthermore, despite the greater customization capacity of open-source models (LLaMa, Mistral, etc.), such as through fine-tuning, these models require technical knowledge and minimal infrastructure, which may compromise the study's replicability. Thus, this point also motivated the choice of ChatGPT, Gemini, and DeepSeek, which provide online interfaces with easily accessible chatbots. Furthermore, for greater accuracy and analysis of the results, it was decided to use two equivalent versions in each LLM: the advanced standard (usually via subscription) and the one emphasizing logic/reasoning. Thus, the following versions of the models were used: 4o and o3-mini-high from ChatGPT, Gemini Advanced 2.0 Flash and 2.0 Pro Experimental from Gemini, and DeepSeek-V3 and DeepThink (R1) from DeepSeek.

Question definition: it was conducted by three GLA experts who (i) have more than six years of experience with educational games, (ii) have been investigating the field of GLA for more than two years, and (iii) two are professors with a PhD, with an emphasis on Informatics in Education and Computer Education. Thus, it was defined that, in addition to analyzing the capture structures generated, it would be necessary to understand how the models interpret theoretical GLA questions from their knowledge bases – which could impact the structures. Therefore, theoretical questions were also listed to analyze the responses of the models, of which: (i) "What is GLA?"; (ii) "How to implement GLA in an educational game?"; (iii) "Who are the main stakeholders of GLA?"; (iv) "What are the challenges of the GLA area?"; (v) "What is GLBoard?"; (vi) "What is the GLBoard capture template?"; (vii) "What is X-API-SG?"; (viii) "In GLA, what is data modeling?". Regarding the activity requested from the model, it was decided to generate capture structures in the GLBoard template. The models are provided with the template, the data labels and their explanations, and the information

from the educational game "Tricô Numérico" (Numeric Knitting). The game's objective is to support the learning of basic mathematical operations. The player must hit zombies that carry the numbers or symbols needed to complete the incomplete mathematical expressions presented correctly. The game was chosen for its simple mechanics, making generating and analyzing GLA structures easier. Then, the models are instructed to fill in the data template based on the game information to help identify evidence of learning from the player's path. The advanced versions of the LLMs are responsible for dealing with GLA issues, while the versions focused on logic and reasoning are accountable for generating capture structures in zero, one, and few-shot learning conditions (Figure 2).

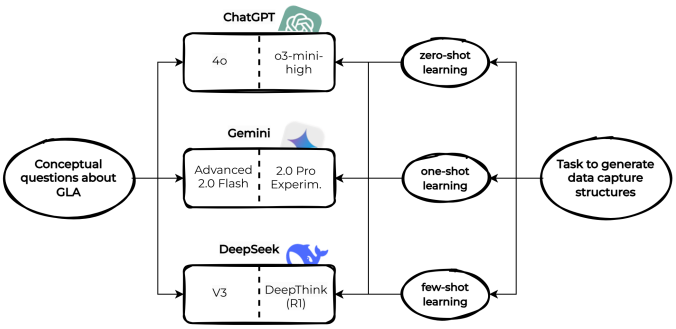
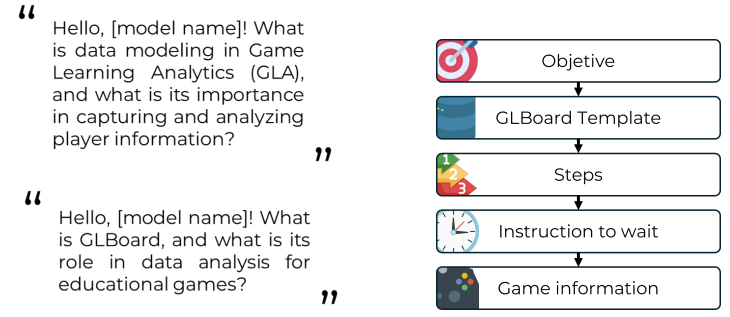


Figure 2. Models of the LLMs used in the experiment.

Prompt construction: the next step involved designing the prompts (instructions) to be sent to the selected LLMs based on previously defined items. For a more in-depth investigation of the models, in the task of generating data capture structures, we chose to use In-context learning in its three conditions: zero, one, and few-shot learning. Thus, eleven prompts were developed: eight for the theoretical questions and three for generating the capture structures (one for each ICL condition). The person responsible for constructing this prompt was one of the experts from the previous step who, in addition to his experience with GLA, has already conducted experiments on prompt engineering and LLMs for this domain.

Refinement with meta-prompting: as a measure to ensure more accurate prompts, we chose to use the meta-prompting technique to refine them. This approach belongs to the field of automatic prompt engineering, whose objective is to guide an LLM to inspect a prompt, provide feedback and update it, minimizing manual efforts in this process [Ye et al. 2023]. The model selected to use the meta-prompt was the "GLA Specialist", a customizable version of ChatGPT [Honda et al. 2024], which was chosen because its knowledge base aggregates the main GLA researches from the literature, in addition to the documentation of the GLBoard system with the detailed explanation of the data capture template. In this scenario, the "GLA Specialist" was used to analyze and propose adjustments to the eleven prompts previously developed. In view of this, the prompts were adjusted, with the main changes being related to fluidity (less formal texts), logical sequence (item enumeration), reduction of ambiguity (avoiding very generic questions), organization (highlighting important points) and standardization of JSON (greater readability). Figure 3(a) contains two prompts that were refined and Figure 3(b) shows how the prompts for generating the data capture structures were structured, containing: (i) Objective – informing the context and objective of the task; (ii) GLBoard

template – providing the capture template with explanatory comments for each field; (iii) Steps – step by step that the LLM should conduct, with emphasis on some aspects; (iv) Instruction to wait – informing the model to wait for the game data that would be sent; and (v) Game information – to be sent in a second prompt. The full list of prompts can be found at the link[1].



(a) Prompts refined through meta-prompting.

(b) Prompt structure for generating capture structures.

Figure 3. Refined prompts and prompt structuring, respectively.

Data collection: involves sending prompts to the models. For each question/activity, a new chat was used to prevent the model from “learning” from the answer to the previous question. As mentioned previously, in the activities to generate the capture structures, it was decided to send two prompts: (i) containing the objective of the activity, the GLBoard capture template, the steps and instructions to wait, along with the examples – in cases of one- or few-shot learning; and (ii) the data of the educational game whose structures will be based. This strategy was adopted to avoid long prompts and not compromise the model’s results. In the one-shot learning condition, the example provided consisted of a completed data capture template related to an educational game to practice the shortest path. This same example was used in few-shot learning and another associated with an educational game focused on basic math operations (addition and subtraction). Both were chosen due to their simple mechanics, which facilitate the structure’s design. The generated responses were stored in an online spreadsheet, obtaining 24 responses to the questions (eight from each model) and nine capture structures (three from each model, in the one-, zero-, and few-shot learning conditions).

Evaluation questionnaire construction: an online form via Google Forms was created to evaluate the models’ responses to the theoretical questions. Each question was divided into sections containing the model’s responses (anonymized to avoid bias), quantitative questions on a Likert-5 scale to evaluate them individually, a question to select which model obtained the most appropriate response, and a text field for the justification. In addition, a section with qualitative questions was included to evaluate the models’ reactions in general, identify if anything caught the attention, and insert additional comments (optional). Regarding the evaluation of data capture structures, no instrument in the literature performs this type of analysis because (i) data modeling is not described in detail in GLA studies, and (ii) the structures are in a format specific to

[1]List of full Prompts on Google Drive (clickable text).

the GLBoard model. As an alternative, GLA experts designed the "Player Level Up": an evaluation model to assign scores to data capture structures in the GLBoard template, which consists of three dimensions (coherence/completeness, redundancy, and evidence of evolution) with a total of seven questions on a Likert-5 scale and an optional field for comments – available in Table 1. Therefore, the nine capture structures generated by the models and the "Player Level Up" questions were inserted in another Google Form.

Table 1. Evaluation model for GLA structures in the GLBoard standard.

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

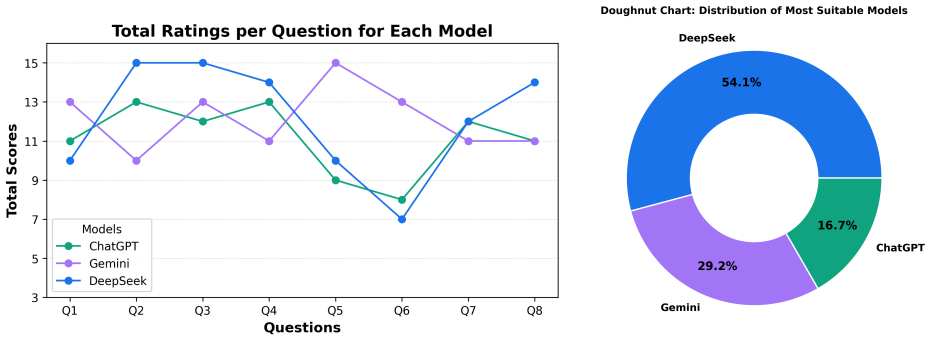
Data analysis: to evaluate the questions and structures generated, three GLA experts were invited – different from those who defined the questions in the previous stage. The experts had years of experience with educational games, most of whom had already implemented GLA techniques. Thus, with the consent that the data would be used only for research purposes, the experts filled out both forms, resulting in 87 quantitative questions evaluated by each expert: 24 of the theoretical questions on GLA and 63 of the evaluations of the capture structures. The Google Forms functionality was used to export the responses to .xlsx (Excel) format and send them to Google Colab to generate histograms and boxplots, aiming to compare the evaluations of the LLMs' responses.

4. Results and discussions

Based on the evaluations of GLA experts for the models' response evaluation forms, graphs were constructed to facilitate the visualization and understanding of the results. Figure 4(a) presents the results of the three models in a line graph, where the X-axis represents each evaluated question (Q1 to Q8) and the Y-axis contains the sum of the experts' scores for each question. Figure 4(b) has a donut graph that illustrates the percentage of models the experts judged to have answered each question most adequately.

Approximate evaluations of all models can be seen, suggesting that their responses are similar. However, at some points, DeepSeek and Gemini obtained excellent results, while ChatGPT never proved superior to the others. The sum of the experts' scores for the responses of DeepSeek and Gemini were the same (97); however, the graph in Figure 4(b) indicates the experts' preference for the results generated by DeepSeek (54.1%).

Regarding the individual analysis of the questions: **Q1:** the three models provided adequate answers to the definition of GLA, emphasizing Gemini, which provides more complete information and examples from the experts' perspective. For this question, the answers were requested to be concise and objective, which ChatGPT and DeepSeek met. Despite Gemini's good answer, it did not meet the objectivity aspect. **Q2:** DeepSeek received the highest score, as it provided more steps and was more detailed in its descriptions for implementing GLA. ChatGPT and Gemini also presented adequate answers, however, with fewer steps. **Q3:** The three models precisely defined the



(a) Sum of Question Ratings per Model.

(b) Distribution of the most suitable models.

Figure 4. Line and donut charts, respectively, of the evaluations.

Stakeholders. DeepSeek presented ethical aspects, grouping information and providing more examples. **Q4:** Similar to Q4, the models obtained similar answers regarding challenges, but DeepSeek presented more complete results. **Q5:** Regarding GLBoard, Gemini was the only one that returned correct answers about the model. However, this is because it has functionalities for web access, while in DeepSeek this function must be activated. Therefore, Gemini's advantage was in this aspect. ChatGPT, on the other hand, despite natively having web access, did not provide coherent results. **Q6:** Regarding the GLBoard template, Gemini, with web access, provided more precise answers about the data capture structure, while the answers of the other models were unsatisfactory. **Q7:** Regarding xAP-SG, there was a more significant divergence among the experts; while one portion indicated that ChatGPT provided a complete description with specific examples, and another commented that only DeepSeek understood the data architecture. **Q8:** Regarding data modeling, the responses were similar for the models, with an emphasis on DeepSeek, which included more detailed responses and an example.

In general, the experts found the models' responses to be interesting and detailed. In some cases, more extensive responses were more appropriate, including examples of use, but in others, they were more generic. Among the points that caught the experts' attention were: (i) the inclusion of solutions for educators to visualize the data collected by GLA; (ii) the models with more extensive responses, in the majority, understood the content more clearly; and (iii) a particular case of a model that returned a response that was different from what was requested. Suggestions include providing information to the models about the technologies and requirements for practically using xAPI-SG and GLBoard. Therefore, the models performed similarly for the theoretical questions, emphasizing DeepSeek, which provided more precise and detailed responses.

Regarding the capture structures generated by the models in the GLBoard template, their evaluation was carried out by filling out a form based on seven questions according to the "Player Level Up" model. Figure 5 presents a boxplot that provides an overview of the evaluations of each model and their respective learning conditions (zero-shot, one-shot, and few-shot). The scores attributed by the experts for each question were added, obtaining values ranging from seven (all questions evaluated with a minimum score) to 35 (evaluation of all questions with a maximum score).

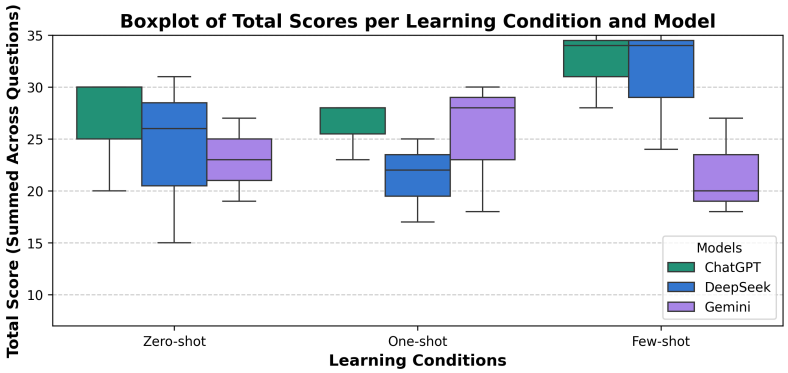


Figure 5. Boxplot chart providing an overview of the evaluations.

As illustrated in the boxplot, models in few-shot learning conditions presented better results than in zero- and one-shot — except Gemini, corroborating studies that indicate examples enhance responses [Brown et al. 2020, Mann et al. 2020]. In general, evaluations in zero- and one-shot were similar, with a subtle advantage for ChatGPT — in one-shot, it showed smaller data dispersion than Gemini. The best evaluated model was ChatGPT, in both zero- and few-shot. DeepSeek also matched in few-shot, with similar results and slightly higher dispersion. In contrast, Gemini had unsatisfactory results in zero-shot and, unexpectedly, in few-shot, regarding capture structure generation. Table 2 provides a more detailed analysis, with model, condition, mean, median, standard deviation (SD), and minimum/maximum values. ChatGPT and DeepSeek, in few-shot, obtained maximum scores, highest means (32.33 and 31), and medians (34). DeepSeek was slightly behind due to a lower minimum, increasing its SD. Gemini in few-shot had one of the weakest means (21.67), the lowest median (20), and one of the lowest minimum (18) and maximum (27) values, scoring worse than in zero-shot.

Table 2. Descriptive Statistics of Evaluations by Model and Learning Condition

Model	Condition	Mean	Median	SD	Min	Max
ChatGPT	Few-shot	32.33	34.0	3.79	28	35
ChatGPT	One-shot	26.33	28.0	2.89	23	28
ChatGPT	Zero-shot	26.67	30.0	5.77	20	30
DeepSeek	Few-shot	31.00	34.0	6.08	24	35
DeepSeek	One-shot	21.33	22.0	4.04	17	25
DeepSeek	Zero-shot	24.00	26.0	8.19	15	31
Gemini	Few-shot	21.67	20.0	4.73	18	27
Gemini	One-shot	25.33	28.0	6.43	18	30
Gemini	Zero-shot	23.00	23.0	4.00	19	27

Regarding the evaluation of the questions, the models showed good results – especially in few-shot learning. The questions that received the lowest scores were Q6 and Q7 by the Gemini and DeepSeek models – mostly with zero-shot and one-shot conditions. These questions deal with essential variables not considered in the structures and temporal analysis of the player's path. This suggests difficulties for the models proposing data for collecting an educational game, such as timestamps. On the other hand, in Q1, Q2, and Q5, variable names, their relationships with game mechanics, and the absence of duplicate variables were the best-rated questions, mainly by the ChatGPT and DeepSeek models in few-shot learning. As for Q3 and Q4 about appropriate data types and variables belonging

to the $path\_player$ field, the results were positive and similar, except for Gemini, which was average in proposing variables related to the player's path.

Limitations and threats to the validity of the study include: (i) GLA experts belong to the same research group, which may generate bias; (ii) the capture structures are from an educational game; (iii) the perception of the experts, not considering the students; and (iv) the hallucination of the LLMs, which may generate inaccurate/incoherent results [IBM 2023]. Although the experts are from the same research group, they were selected based on their practicality and experience in GLA, which leads us to believe that they did not influence the results. An analysis with more educational games and the perception of the students—although important—would increase the complexity of the study, whose objective was to present introductory research on the performance of the models with GLA activities. Therefore, they will be considered in future research. Regarding the hallucination of the models: when asking about the GLBoard without activating the web search, which will be studied later.

5. Conclusions

Given the contributions of GAI with the advances of LLMs in several domains, such as GLA, this paper consisted of an empirical study guided by the research question: "how do ChatGPT, Gemini, and DeepSeek perform in specific GLA activities?". To this end, theoretical questions were developed to send to the models and activities envolving the generation of capture structures in the GLBoard template and the construction of prompts. These instructions were refined and sent to the models using meta-prompting, collecting the data, and storing it in a spreadsheet. Three GLA experts were invited to evaluate the responses of the models, filling out two forms: one to assess the theoretical questions through Likert-5 scale scores and another based on the "Player Level Up" model, related to the evaluation of the capture structures in the GLBoard JSON standard.

In response to the research question, the models performed similarly in GLA activities but in different ways. In theoretical questions, both DeepSeek and Gemini were the best evaluated, with experts judging DeepSeek's responses to be more adequate. The few-shot learning condition proved to be the most efficient for capture structures compared to zero-shot and one-shot. The best-evaluated model was ChatGPT, with DeepSeek slightly behind. Gemini presented the worst results in few-shot learning. Regarding of evaluation of the structures, the least well-evaluated were about variables not considered in the structures and the absence of time records, revealing the difficulty of the models in appropriating and understanding the educational game. The models obtained satisfactory results regarding variable names, relations with the game mechanics, and duplicated variables, indicating coherence in the structures.

The contributions of this study indicate that (i) models obtain different results for the same activities, such as in GLA; (ii) DeepSeek, with less than two months of use, obtains results comparable to ChatGPT and Gemini, considered one of the best LLMs; (iii) few-shot learning conditions enhance the models' responses; and (iv) there is potential for LLMs to assist in the understanding and implementation of GLA techniques, such as in the construction of capture structures. Future studies include (i) analyzing more issues with LLMs, (ii) providing contexts about other educational games, and (iii) inviting more experts for a more accurate evaluation.

6. Acknowledgment

In this study, Generative AI (GenAI) was used through Chat-GPT 4o from OpenAI to generate the codes for graphs in Overleaf, aiming to help minimize time and effort in constructing these representations.

This study was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) - Finance Code 001 and Conselho Nacional de Desenvolvimento Cientifico e Tecnologico - Brasil - CNPq (Process 303443/2023-5). This work was partially funded by the Amazonas State Research Support Foundation – FAPEAM – through the PDPG/CAPES/FAPEAM project.

References

Akter, S. N., Yu, Z., Muhamed, A., Ou, T., Bäuerle, A., Cabrera, Á. A., Dholakia, K., Xiong, C., e Neubig, G. (2023). An in-depth look at gemini's language abilities. *arXiv preprint arXiv:2312.11444*.

AlBadarin, Y., Tukiainen, M., Saqr, M., e Pope, N. (2023). A systematic literature review of empirical research on chatgpt in education. *Available at SSRN 4562771*.

Alonso-Fernandez, C., Calvo, A., Freire, M., Martinez-Ortiz, I., e Fernandez-Manjon, B. (2017). Systematizing game learning analytics for serious games. In *2017 IEEE global engineering education conference (EDUCON)*, pages 1111–1118. IEEE.

Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., e 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.

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

BBC News (2025). DeepSeek: The Chinese AI app that has the world talking. *BBC News*. Accessed: 2025-02-05.

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.

Chan, A., Kharkar, A., Moghaddam, R. Z., Mohylevskyy, Y., Helyar, A., Kamal, E., Elkamhawy, M., e Sundaresan, N. (2023). Transformer-based vulnerability detection in code at edittime: Zero-shot, few-shot, or fine-tuning? *arXiv preprint arXiv:2306.01754*.

DemandSage (2025). Chatgpt statistics 2025: Users, traffic, revenue more. Accessed: 2025-02-17.

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

Honda, F., Pires, F., Pessoa, M., e 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.

IBM (2023). Ai hallucinations. https://www.ibm.com/topics/ai-hallucinations. Accessed: 2025-04-20.

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

Jovanovic, M. e Campbell, M. (2022). Generative artificial intelligence: Trends and prospects. *Computer*, 55(10):107–112.

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

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., et al. (2023b). Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017.

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.

Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Nexocode (2023). Customizing large language models: A comprehensive guide. Accessed: 2025-02-17.

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

Saveski, G. L., Westera, W., Yuan, L., Hollins, P., Manjón, B. F., Ger, P. M., e Stefanov, K. (2016). What serious game studios want from ict research: identifying developers' needs. In *Games and Learning Alliance: 4th International Conference, GALA 2015, Rome, Italy, December 9-11, 2015, Revised Selected Papers 4*, pages 32–41. Springer.

Sengar, S. S., Hasan, A. B., Kumar, S., e Carroll, F. (2024). Generative artificial intelligence: a systematic review and applications. *Multimedia Tools and Applications*, pages 1–40.

SEO.ai (2025). Deepseek users, downloads, and growth. Accessed: 2025-02-17.

Silva, D., Pires, F., Melo, R., e 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.

Singh, H. e Singh, A. (2023). Chatgpt: Systematic review, applications, and agenda for multidisciplinary research. *Journal of Chinese economic and business studies*, 21(2):193–212.

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

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., e Yang, Q. (2023). Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.