BİTLİS EREN ÜNİVERSİTESİ
FEN BİLİMLERİ DERGİSİ

ISSN: 2147-3129 / e-ISSN: 2147-3188

Article Type	: Research Article	**Year**	: 2025
Received	: March 25, 2025	**Volume**	: 14
Revised	: July 2, 2025	**Issue**	: 3
Accepted	: July 8, 2025	**Pages**	: 1469-1486
DOI	: 10.17798/bitlisfen.1664312		

GDD GENERATION FOR HYPER-CASUAL GAMES USING LARGE LANGUAGE MODELS: A COMPARATIVE EVALUATION

Muhammet Emin AYDINALP [1, *], Buket DOĞAN [2], Abdullah BAL[2]

[1] *İstanbul Sabahattin Zaim University, Computer Engineering Department, İstanbul, Türkiye*

[2] *Marmara University, Computer Engineering Department, İstanbul, Türkiye*

**** Corresponding Author:*** *muhammet.aydinalp@izu.edu.tr*

ABSTRACT

Game Design Documentation (GDD) is a critical document that includes the design and mechanical details of the game to be developed. These documents create a common understanding among team members by including details such as the game's progress, story, and design features. In order for the game development process to proceed and be completed healthily, these documents must be prepared in a high-quality, clear, and detailed manner. However, the creation of this documentation is a time-consuming and error-prone process. Especially in game genres that require rapid prototyping, incomplete or insufficient GDDs can cause delays in the project process. This study was conducted to examine the effectiveness of LLMs in GDD production. The hyper-casual game Pool Wars was selected as a reference, and for this example game, the GDD created by a human expert and the GDD produced by ChatGPT-4 using various prompt methods were evaluated by four experts in the field according to eight different criteria using a five-point Likert scale. In addition to structural and creative aspects, visual elements were also included in the evaluation process. ImageFX, developed by Google, was used to add visual content to the GDD created by ChatGPT-4. As a result, it was seen that LLMs were more successful in many criteria in GDD production. As a result of the scoring made by an academician and three experts from the sector, GDD created by LLM received an overall average score of 4.71 out of 5, while GDD prepared by human expert received 3.29 points. GDD produced by LLM showed a clear superiority especially in terms of understandability and level of detail. However, it showed a similar performance to human expert in terms of creativity and visual content and it was observed that there was room for improvement in these areas.

Keywords: Large language models (LLM), Game design documentation (GDD), ChatGPT-4, Hyper-Casual games, Prompt engineering, ImageFX.

1 INTRODUCTION

Game design documentation (GDD) is a very important part of the game production pipeline [1], [2]. These documents guide team members through each stage of the production cycle by describing in detail the concept, mechanics, story, and other main components of the game [3]. Today, the historical timeline of GDD has evolved from simple text documents in the early 1990s to comprehensive documentation that integrates with multimedia content, software project management tools, and version control systems today [4]. In interviews with game development industry employees, one of the most common complaints is the difficulty and inconsistency of the GDD preparation process. Especially in hyper-casual [5] (simple mechanics, fast playable and appealing to a wide audience) game studios, teams working under rapid production pressure have difficulty preparing comprehensive and consistent GDDs, which leads to serious disruptions in project management [6], [7].

In recent years, Large Language Models (LLMs) [8] have revolutionized the fields of artificial intelligence and natural language processing. LLM is a pre-trained language model based on a vast amount of textual data using transformers and deep learning techniques [9]. These models are used in many areas such as text generation, translation, speech recognition, text analysis and classification [10]. Recently, LLMs have gained the ability to generate audio, visual and video from text and very wide content has been produced using these features [11] - [13]. The use of LLMs has become widespread in various sectors such as education, cinema, medicine, cybersecurity and game development and has made significant contributions to these fields [14]-[16].

One of the studies conducted using LLMs is the creation of Software Requirements documents by LLMs, crosscheck with human experts and performance analysis. In the study of Krishna et al., a They assess the performance of GPT-4 and CodeLlama in drafting an SRS for a university club management system and compare it against human benchmarks using eight distinct criteria [17]. The results of the study showed that LLMs can produce documents that are equivalent in quality to SRS documents produced by entry-level software engineers. Similarly, Lubos et al.'s study introduce and assess the capabilities of a Large Language Model (LLM) to evaluate the quality characteristics of software requirements according to the ISO 29148 standard [18]. They show how an LLM can assess requirements, explain its decision-making process, and examine its capacity to propose improved versions of requirements. In a study conducted by Lee et al. [19], GPT-3 and LLaMA-2 models are trained to answer questions

related to technical specification documents. The results show that the fine-tuned LLaMA2 model generally outperforms the fine-tuned GPT-3 model in terms of accuracy, reliability, and conciseness of responses. it has been proven by various studies that open-source models can also exhibit high performance when fine-tuned for a specific domain [20]. This shows that secret documents and information can be processed securely through LLMs without being shared with external systems [21].

In this study, a hypercasual game called "Pool Wars" was selected. For this game, there is a GDD written by an expert who has worked as a game designer in the game industry for more than 5 years and has industry experience, and a GDD was created for the same game with ChatGPT-4 [22] using various prompting methods such as giving clear instructions, step-by-step development, supporting with examples and role-playing prompting. The visual part of the created GDD was completed with ImageFX [23]. As a result, the GDD generated by LLM and the human-generated GDD were evaluated by experts in the field and compared by scoring according to a five-point Likert scale. In the study, the evaluation criteria used for both GDD and SRS documents were scored according to the following criteria: completeness, understandability, consistency, creativity and innovation, applicability, suitability for the target audience, level of detail and user-orientedness.

This study aims to display the competence, quality and consistency, time and efficiency advantages and potential contributions of LLMs in game design documentation production. It is also aimed to determine the capacity and weaknesses of LLMs and make recommendations for future developments.

The contributions of this article are as follows:

• Adequacy of LLMs in GDD Production: This study demonstrates how effective and adequate LLMs are in producing game design documentation when compared to GDDs created by human experts, and also evaluates the attribute and consistency of GDDs produced by LLMs and specifies how useful these documents are in the game development process.

• Recommendations for Development: It identifies the strengths and weaknesses of LLMs in GDD production and makes suggestions for future improving of these models.

• Application of Evaluation Criteria: The evaluation criteria used in the study are valid for both GDD and SRS documents, and demonstrate how these criteria can be used to objectively measure the performance of LLMs.

2 MATERIAL AND METHOD

In this study, the effectiveness of LLMs in the GDD creation process was investigated. For a hypercasual game called "Pool Wars", a GDD prepared by human experts was taken as a reference. Then, a similar GDD was created using the ChatGPT-4 model. In this process, various prompt engineering methods such as providing clear instructions, step-by-step guidance, supporting with examples and role-playing techniques were applied. The study also incorporated visual elements generated through ImageFX to complement the textual content from the language model. Four game design and development specialists assessed both GDDs using eight evaluation criteria, with ratings assigned on a five-point Likert scale. The evaluation methodology included both quantitative and qualitative approaches, enabling a comprehensive comparison that identified the capabilities and limitations of large language models in game documentation tasks.

2.1 Data Preparation

In this study, a hypercasual game called ‘Pool Wars’ was selected for the target GDD creation process. Hypercasual games are one of the game genres where the GDD process is critical due to their simple mechanics, requiring rapid prototyping and appealing to a wide range of players. Therefore, they constitute a suitable example to evaluate the effectiveness of LLMs in the GDD creation process. A GDD prepared by human experts for this game was used as the base document. This GDD was prepared for a game previously developed by an expert Game Designer in a professional game studio that produces hypercasual games in the industry and was used as an example GDD produced by a human expert in this study. In order to compare with the human-authored GDD, a second GDD was generated using GPT-4, a large language model developed by OpenAI. The resulting document was later evaluated by field experts alongside the human-authored GDD. The full version of the LLM-generated GDD has been made publicly accessible via Zenodo [24]. The prompts used in the creation of the LLM-based GDD were structured to ensure clarity, consistency, and detail throughout the document. These prompt [25] strategies are explained below.

2.1.1 Step by Step Instructions Prompt

Since LLMs have limited tokens and context windows at a time, we divided the document we wanted to create into pieces and proceeded step by step. Game design documents

already consist of certain sections. We gave special prompts for these sections and combined the created texts. First, we gave a prompt [26] explaining which headings a GDD should contain and that we wanted to produce these headings step by step. Example prompt: “A game design document consists of headings such as game summary, basic mechanics, level system, design elements, game economy. We will create a game design document where we will fill in these headings for our hyper casual game called Pool Wars, a pool game.” Then, we gave a prompt talking about what kind of game I wanted to make and asked for the above-mentioned headings one by one to create our document. Sample prompt: “The story of the stickman character who embarks on an adventure on the seabed in a pool environment suitable for the summer theme. The only main character, the warrior stickman, tries to defeat the enemies on the battlefield by making the necessary improvements in the safe area allocated to him in the pool. With each enemy he destroys, he continues his fight in new pools by collecting the necessary equipment, loot and improvements. Before discovering new fronts, he faces a dangerous end-of-level creature. One of the goals of our main character is to continue on his way by defeating the boss character. Can you create the basic mechanics for the game given above in detail?”

The above prompt was repeated for each main title and the created texts were combined [27].

2.1.2 Example Supported Prompt

In the GDD creation process of the model, it was aimed to guide the output and increase its quality by using GDD examples of similar games [28]. In this method, sample GDD content was provided in a format that the model could understand and process. These examples provided guidance on what kind of output the model should produce, both structurally and content-wise. Especially in sections such as game mechanics, level design and storytelling, sample GDDs helped the model produce more comprehensive and detailed output. In this way, the model was enabled to learn from concrete examples and be guided more specifically, instead of just general instructions.

2.1.3 Role Playing Prompts

In the GDD creation process of ChatGPT-4, role-playing prompts were used to ensure that the model took on a certain role and produced outputs appropriate to this role [29]. In this method, the model was given a prompt such as "You are a game designer. You are knowledgeable about hypercasual games. Focus on the target audience and suggest intuitive

gameplay mechanics while creating the design document of this game" and was asked to take on the role of a game designer and create a GDD from this perspective.

The role-playing prompts aimed to make the model think from the perspective of a game designer and create a GDD that was appropriate for the target audience, creatively and applicable. This method allowed the model to not only provide information but also to think like a game designer and produce a more original and user-focused GDD.

For example, when the model took on the role of "game designer", it suggested simpler and more intuitive gameplay mechanics considering the target audience of the game (e.g., young players). It also made suggestions to make the visual and audio design of the game appealing to the target audience. Role-play prompts helped the model create a more comprehensive and realistic GDD.

The visual elements of the GDD produced by ChatGPT were created using ImageFX, a text-to-image conversion tool. This tool transformed text-based descriptions into visuals, making the GDD more understandable and engaging. ImageFX was used to create various visual elements such as in-game characters, environments, objects, and the user interface.

The GDDs produced by both human experts and LLM were organized as comprehensive documents covering sections such as game mechanics, story, visual style, and user interface. In these sections, the basic elements of the game were explained in detail and supported by visuals. For example, in the game mechanics section, elements such as the basic rules of the game, the gameplay loop, and control mechanisms were explained. In the story section, information was given about the game's story, characters, and world. In the visual style section, elements such as the game's general art style, color palette, and character designs were discussed. In the user interface section, game menus, indicators, and other interface elements were described in detail.

These comprehensive GDDs were designed to serve as a guide for all team members during the game development process. GDDs were created to clearly document the design and content of the game and create a common understanding among all stakeholders.

2.2 Evaluation Criteria

The GDDs created were evaluated in detail by experts in the field. During the evaluation process, communication with the experts was provided through an Excel file containing the evaluation criteria for both GDDs. The experts were sent in PDF format without specifying which source produced both GDDs and were asked to make an objective evaluation. In the

Excel file sent to them, areas were provided where they could score each GDD separately based on the eight criteria specified. At the end of the evaluation process, the experts filled in their scores in the Excel file and shared their qualitative observations by providing additional written comments. Thus, both quantitative (scoring data) and qualitative (comments and opinions) data were collected and analyzed. The evaluation criteria used in this study were determined by considering previous studies in the literature on the evaluation of game design documentation [30].

These studies emphasize that game documentation should be considered from various aspects such as comprehensiveness, understandability, consistency, creativity, applicability, level of detail, user-centeredness, and visual adequacy [31]. These criteria include the elements that are important for determining the quality and effectiveness of a GDD. The experts who conducted the evaluation used a five-point Likert scale [32]. In this scale, which is explained in detail in Table 1, 1 represents the lowest score and 5 represents the highest score.

Table 1. Likert Scale for GDD Performance

Score	Statement	Description
1	Strongly Disagree	The performance of GDD is extremely inadequate.
2	Disagree	The performance of GDD is inadequate
3	Neutral	The performance of GDD is neither good nor bad.
4	Agree	The performance of GDD is good.
5	Strongly Agree	The performance of GDD is excellent.

Table 2 provides a detailed list and explanation of all the criteria used in the evaluation of GDDs. Experts carefully examined each criterion and scored the GDDs on this scale.

Table 2. Evaluation Criteria

Evaluation Criteria	Description
Completeness	Whether the document covers all critical elements.
Clarity	The clarity and ease of understanding of explanations.
Consistency	Logical coherence and integrity between sections.
Creativity & Innovation	The originality and innovative nature of the presented ideas.
Practicality	The document's practicality and suitability for the target audience.
Level of Detail	The depth and precision of the provided explanations.
User-Centric Approach	Alignment with the needs and expectations of the end user.
Visual Adequacy	The quality and clarity of visual elements.

In this study, four professionals with backgrounds in game design and development took part in the evaluation of the GDDs. Among them, three are actively working in the industry as game developers, each with a minimum of three years of experience. The fourth expert has been involved in academic research on game design for over a decade. Although their areas of expertise differ, all experts assessed the documents based on the same set of criteria without any distinction in responsibility. During the evaluations, the GDDs were examined in terms of several important aspects: how complete and understandable they were, whether the content maintained consistency throughout, how original and creative the ideas appeared, the practical value of the documents, the level of detail they provided, how well they addressed user needs, and the adequacy of the visual elements included. Having experts from both industry and academia enriched the process by bringing in multiple perspectives. This diversity helped create a more balanced and comprehensive assessment overall.

To avoid bias, each expert reviewed the GDDs independently. They were not informed which document was written by a human and which was created with the help of a language model. This anonymity helped ensure that the evaluations were fair and unbiased. Once all individual evaluations were completed, we collected and analyzed the results together. We conducted a detailed comparative analysis to determine the performance differences between the human-written and LLM-generated GDDs. We started by calculating the average score each GDD received under each evaluation criterion. This step allowed us to see how well each document performed in certain aspects. We also calculated the differences between the means in order to see which GDD performed better under which criterion. We calculated the standard deviation for each criterion to observe how consistent the scores were across raters. Beyond the numerical results, we also took into account the written feedback of the experts. Each participant shared their personal observations about the strengths and weaknesses of both GDDs. These insights provided important context for the scores and offered valuable information for improving the quality of AI-assisted game design documentation in future work.

2.3 Tools and Technologies Used

In this study, various tools and technologies were used to create the textual and visual content of GDD and to visualize and analyze the data obtained after the comparison.

2.3.1 ChatGPT-4

In this study, we used the free limited version of GPT-4 Plus [33]. This version is directly accessible through OpenAI's web interface. However, this version can only be used for a certain period of time, with an automatic downgrade to a lower version when the time is up. As a result, some practical challenges related to the use of the free version were encountered during the document production process. For instance, after a certain amount of usage, the system occasionally prompted us to start a new chat or required a waiting period before continuing. Consequently, the content generation process had to be split across different chat sessions or resumed at intervals. These interruptions required manually merging content from multiple sessions and maintaining contextual consistency between sections.

However, for the purpose of this study, this version was deliberately chosen to assess how large language models can be utilized by general users for GDD production, without the need for professional tools or special access. The sections of the GDD were developed step by step using the prompt techniques described above.

2.3.2 ImageFX

The free version of ImageFX was used to generate the visual components of the GDD created by ChatGPT [34]. ImageFX is an AI-powered tool capable of producing images based on text-based descriptions. Using this tool, various visual elements such as stickmen, swimming pools, collectible items, and water guns—as well as components of the user interface—were added to the GDD to enrich its visual content.

2.3.3 Microsoft Excel and Python

Microsoft Excel and Python were used to organize, analyze and visualize the evaluation data. The scores given by the experts were organized in Excel and analyzed using Python. Statistical calculations such as mean, median and standard deviation were made using data analysis libraries in Python. In addition, the visualization of the data was provided using matplotlib, one of the graphics libraries in Python. In this way, we made our results concrete and analyzed them.

3 RESULTS AND DISCUSSION

In this study, a comparative analysis of GDDs created by human experts and LLM was performed. The evaluations made by four different experts were examined based on eight different criteria and the results were analyzed both quantitatively and qualitatively. Table 3 shows the average scores and standard deviations according to the criteria.

Table 3. Average Scores and Standard Deviations by Criteria

Criteria	Human GDD (Avg.)	Human SD	LLM GDD (Avg.)	LLM SD	Difference
Completeness	3.67	0.82	4.67	0.50	+1.00
Clarity	3.33	0.00	5.00	0.00	+1.67
Consistency	3.00	0.58	4.67	0.00	+1.67
Creativity & Innovation	4.00	0.50	4.00	0.00	0.00
Practicality	3.00	0.50	5.00	0.00	+2.00
Level of Detail	2.67	0.00	5.00	0.00	+2.33
User-Centric Approach	3.00	0.50	5.00	0.50	+2.00
Visual Adequacy	3.67	0.00	4.33	0.82	+0.66

As shown in Table 4, the LLM-generated GDD received an overall average score of 4.71, while the human-generated GDD received a score of 3.29. The average scores for each criterion and the differences between them are presented in detail in this table. The LLM GDD showed clear superiority especially in the criteria of understandability, applicability and level of detail.

Table 4. Statistical Summary

Metric	Human GDD	LLM GDD
Mean	3.29	4.71
Standard Deviation	0.36	0.23
Minimum Score	2.67	4.00
Maximum Score	4.00	5.00

The statistical summary presented in Table 4 shows the overall performance metrics of both GDDs in comparison. These data reveal that the LLM GDD is superior not only in mean scores but also in rating consistency. Statistical analysis showed that the LLM GDD had more consistent scores across raters (SD = 0.23). The standard deviation of the human GDD was calculated as 0.36.

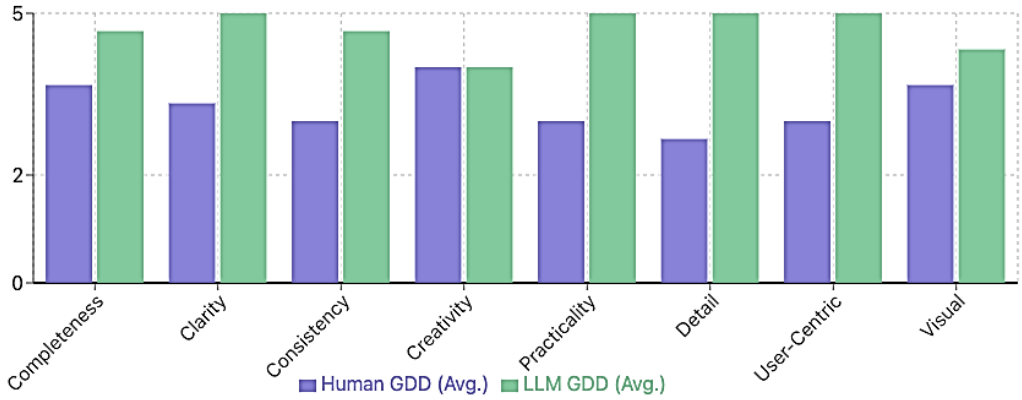

Figure 1. GDD Comparison Bar Chart by Criteria

The bar chart shown in Figure 1 visually presents the comparative performance of GDDs on each evaluation criterion. This chart clearly reveals the large difference observed especially in the level of detail (LLM: 5.00, Human: 2.67).

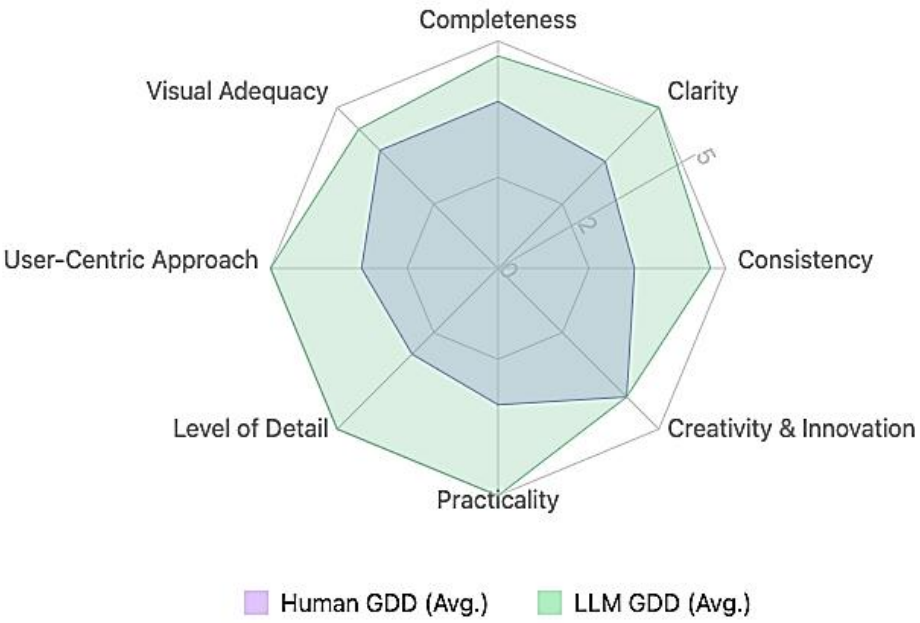

Figure 2. GDD Performance Comparison Radar Chart

The radar chart presented in Figure 2 shows the performance profile of both GDDs holistically. This visualization clearly shows that the LLM GDD consistently performs highly on all evaluated criteria. Of particular note is that in the creativity and innovation criterion, both GDDs perform equally with a score of 4.00.

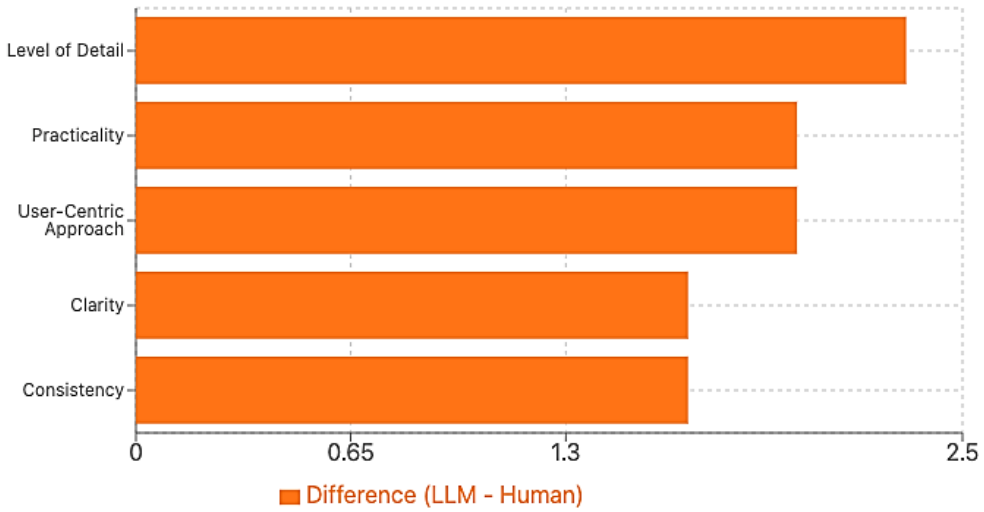

Figure 3. Differences between LLM and Man-Made GDD Criteria

Figure 3 clearly shows the criteria where LLM is significantly superior to human-made GDD. Level of Detail has the highest value with a difference of 2.33, which shows that LLM pays great attention to details when preparing game design documentation. This result indicates that LLM can provide the details missing in human documentation in a clear and explanatory manner.

According to the qualitative observations of the evaluators, it was stated that the GDD produced by LLM exhibited a detailed approach especially in the design of the levels. It was stated that the visual theme, gameplay mechanics, difficulty levels and target scores of each level were defined in detail. It was evaluated that in the GDD produced by LLM, it was clearly explained what kind of experience each level should offer to the player, and this made the design process more effective. It was observed that LLM created a detailed roadmap regarding which tasks should be completed in which weeks until the delivery date of the game. It was stated that this roadmap covers stages such as prototyping, level design, game testing and final editing, and is structured in a way that facilitates project management.

Evaluators believe that this detailed approach shows that LLM is an effective tool not only in content production but also in planning the development process. It was stated that the lack of such details in human-made documentation often leads to delays in project durations, whereas the clarity and openness provided by LLM has the potential to minimize such problems.

Other prominent criteria were Applicability and User Focus, both of which have a difference of 2.00. This shows that LLM has a superior performance in making game documentation practically applicable and taking into account end-user needs. Understandability

and Consistency criteria have a difference of 1.67. This shows that LLM can produce not only detailed but also clear and logically organized documentation.

Figure 4 and Figure 5 below were produced with the ImageFX tool to evaluate the visual adequacy of the GDD created by LLM. GDDs are prepared to provide guiding information for the developer and modeling teams on how to design objects, characters, and environmental elements in the game. In this context, Figure 4 includes various water-themed objects that can be used in the game (e.g. lifebuoys, wave effects, inflatable platforms) and provides a basic visual framework for the general mechanics of the game. Figure 5 shows the interaction of the player character with a stylized water gun and the wave effect in the background. These visuals offer a simple, colorful, and eye-catching visual language in line with the hyper-casual structure of the game; they manage to reflect the concept stated in the GDD in general terms. However, according to expert evaluations, these visual contents have aspects that can be improved in terms of level of detail, use of perspective, and artistic consistency. This situation reveals that LLM-supported GDD production is functional in terms of visual guidance, but it would be useful to support it with professional visual designs for more advanced game projects. Additionally, to access the entire GDD created by LLM and examine all the images, the document shared publicly on Zenodo can be accessed from source [24].

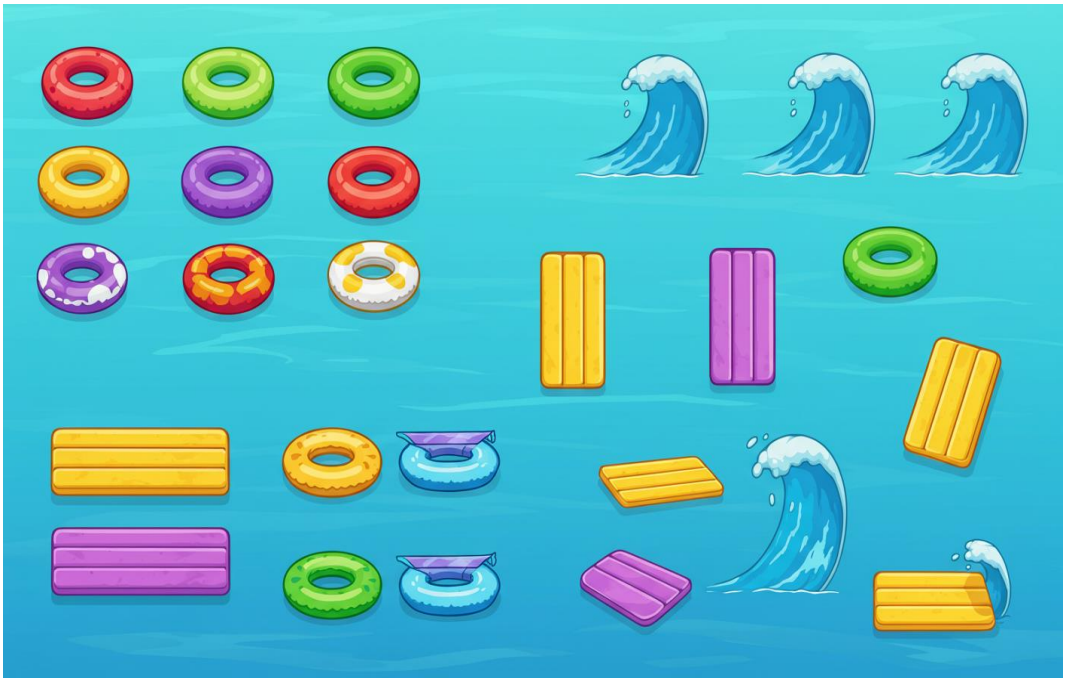

Figure 4. Representative Visuals of In-Game Items and Environmental Assets

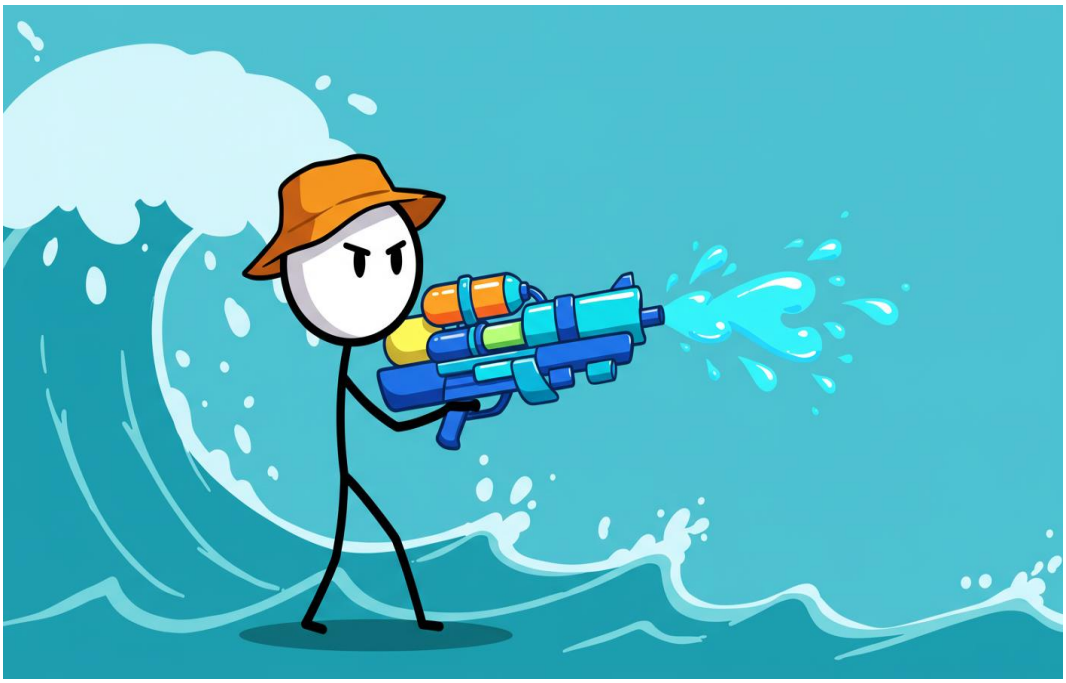

Figure 5. Player Character Interaction with Water Gun and Environmental Elements

Overall, these data suggest that LLM offers strong potential for preparing more structured, detailed, and user-focused documentation than human-made documentation. However, the lower difference in areas such as creativity and visuality suggests that these areas still have room for improvement for LLMs.

Expert feedback has highlighted the strengths and potential impacts of LLM GDD. It has been emphasized that quality issues in GDD writing, especially in the hypercasual game genre, cause serious delays and problems in projects. Experts have stated that insufficient attention is paid to GDD writing in rapid prototyping processes and that the lack of clear definition of game boundaries causes delays in project delivery dates. In this context, the solution offered by LLMs is particularly valuable. Experts have emphasized that the ability of LLMs to create detailed, descriptive, and visually rich GDDs very quickly is an important development for the industry.

Despite the success of LLMs in producing GDDs, there are also some risks. The main ones are that LLMs can add conflicting mechanics and technically impossible features to different parts of GDDs. In addition, producing with exact LLMs can cause a decline in generating new ideas and creativity. In order to eliminate such risks, as suggested in this study, each stage of GDD production should be controlled by human experts and the AI should be guided with the relevant prompts. In other words, it is suggested that GDD production should be done step by step and that each step should be checked and approved before moving on to the next step.

4 CONCLUSION AND SUGGESTIONS

In this study, we explored how effective and adequate LLMs are in generating game design documentation. When we compared the outputs produced by LLMs to those written by human experts, we found that LLMs generally performed better—particularly in areas like completeness, clarity, and level of detail. These findings reflect similar results reported in the work of Krishna et al. [9], where GPT-4 was shown to outperform junior software engineers in tasks related to software requirements documentation. In our case as well, LLMs excelled especially for their ability to produce rich, detailed, and well-structured content that helps guide the design process more effectively than traditional, human-written documents.

The study findings revealed that LLMs can make significant contributions to game development processes in terms of quality, speed and consistency. In this study conducted on the hypercasual game "Pool Wars", it was determined that GDD produced by LLM can eliminate documentation deficiencies encountered in rapid prototyping processes and provide an effective solution to documentation problems in the industry. Expert evaluations have highlighted the contribution of LLM towards standards compliance, structuring and technical details. The results of this study, as emphasized in the study of Lubos et al. [10], show that LLMs are not only a support tool but also a powerful tool that can enable standardization in game design documentation processes.

It should not be forgotten that this success demonstrated by LLM is possible with the guidance of human experts. Otherwise, it should not be ignored that using LLM alone may create problems in terms of originality, consistency and ethics. In addition, in the case of commercial use, issues such as who will own the text produced by LLM and whether there will be copyright problems vary according to the model used. The approaches of different LLMs to this issue are the subject of a different study. In the future, GDD production can be worked on for larger and more comprehensive games using LLMs. Also, how existing LLMs can be improved in terms of visual content production and innovative ideas is a separate study topic.

Conflict of Interest Statement

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The study is complied with research and publication ethics.

Artificial Intelligence (AI) Contribution Statement

Since the subject of this study is to investigate the effectiveness of artificial intelligence (AI) tools in creating game design documents, the GDD mentioned and experimentally used in the manuscript was prepared with ChatGPT-4. In addition, in order to provide visual support for the GDD, the images presented as examples in the manuscript were generated using ImageFX. Furthermore, during the overall writing process of the manuscript, AI tools were utilized for grammar and punctuation corrections.

Contributions of the Authors

Muhammed Emin took part in the creation of game design documents using large language models, the creation of surveys, and the writing of the article. Buket Doğan contributed to the theoretical planning of the article, the creation of surveys, and the academic writing of the methodology section. Abdullah Bal contributed to the implementation and evaluation of the surveys.

REFERENCES

[1] M. G. Salazar, H. A. Mitre, C. L. Olalde, and J. L. G. Sánchez, "Proposal of Game Design Document from software engineering requirements perspective," in *2012 17th International Conference on Computer Games (CGAMES)*, 2012: IEEE, pp. 81-85.

[2] E. Bethke, *Game development and production*. Wordware Publishing, Inc., 2003.

[3] J. Haltsonen, "Guide to Writing a Game Design Document," 2015.

[4] C. Macklin and J. Sharp, *Games, Design and Play: A detailed approach to iterative game design*. Addison-Wesley Professional, 2016.

[5] A. Charoenpruksachat and P. Longani, "Comparative study of usability evaluation methods on a hyper casual game," in *2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering*, 2021: IEEE, pp. 153-156.

[6] П. Івасюк, "Game design document," 2016.

[7] C. M. Kanode and H. M. Haddad, "Software engineering challenges in game development," in *2009 Sixth International Conference on Information Technology: New Generations*, 2009: IEEE, pp. 260-265.

[8] B. Min *et al.*, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys,* vol. 56, no. 2, pp. 1-40, 2023, doi: https://doi.org/10.1145/3605943.

[9] A. Fantechi, S. Gnesi, L. Passaro, and L. Semini, "Inconsistency detection in natural language requirements using chatgpt: a preliminary evaluation," in *2023 IEEE 31st International Requirements Engineering Conference (RE)*, 2023: IEEE, pp. 335-340.

[10] V. Bertram, H. Kausch, E. Kusmenko, H. Nqiri, B. Rumpe, and C. Venhoff, "Leveraging Natural Language Processing for a Consistency Checking Toolchain of Automotive Requirements," in *2023 IEEE 31st International Requirements Engineering Conference (RE)*, 2023: IEEE, pp. 212-222.

[11] Y. C. Gönültaş, "Yapay Zekâ ve Bilimsel Metin Yazımı: Türk Kamu Yönetimi Alanyazınında ChatGPT4. 0 Örneği," *Uluslararası Yönetim Akademisi Dergisi,* vol. 7, no. 3, pp. 827-843, 2024, doi: https://doi.org/10.33712/mana.1578165.

[12] A. Goslen, Y. J. Kim, J. Rowe, and J. Lester, "Llm-based student plan generation for adaptive scaffolding in game-based learning environments," *International Journal of Artificial Intelligence in Education,* pp. 1-26, 2024, doi: https://doi.org/10.1007/s40593-024-00421-1.

[13] M. C. Laupichler, J. F. Rother, I. C. G. Kadow, S. Ahmadi, and T. Raupach, "Large language models in medical education: comparing ChatGPT-to human-generated exam questions," *Academic Medicine,* vol. 99, no. 5, pp. 508-512, 2024, doi: 10.1097/ACM.0000000000005626.

[14] D. Luitel, S. Hassani, and M. Sabetzadeh, "Using language models for enhancing the completeness of natural-language requirements," in *International working conference on requirements engineering: foundation for software quality*, 2023: Springer, pp. 87-104.

[15] G. Agrawal, K. Pal, Y. Deng, H. Liu, and Y.-C. Chen, "CyberQ: Generating Questions and Answers for Cybersecurity Education Using Knowledge Graph-Augmented LLMs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, no. 21, pp. 23164-23172.

[16] J. Pereira, J.-M. López, X. Garmendia, and M. Azanza, "Leveraging Open Source LLMs for Software Engineering Education and Training," in *2024 36th International Conference on Software Engineering Education and Training (CSEE&T)*, 2024: IEEE, pp. 1-10.

[17] M. Krishna, B. Gaur, A. Verma, and P. Jalote, "Using LLMs in Software Requirements Specifications: An Empirical Evaluation," *arXiv preprint arXiv:2404.17842,* 2024.

[18] S. Lubos *et al.*, "Leveraging llms for the quality assurance of software requirements," in *2024 IEEE 32nd International Requirements Engineering Conference (RE)*, 2024: IEEE, pp. 389-397.

[19] J. Lee, W. Jung, and S. Baek, "In-house knowledge management using a large language model: focusing on technical specification documents review," *Applied Sciences,* vol. 14, no. 5, p. 2096, 2024.

[20] Y. Ma, Z. Liu, and O. Kalinli, "Effective Text Adaptation For LLM-Based ASR Through Soft Prompt Fine-Tuning," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024: IEEE, pp. 64-69, doi: 10.1109/SLT61566.2024.10832227.

[21] D. Raj, G. Keren, J. Jia, J. Mahadeokar, and O. Kalinli, "Faster speech-llama inference with multi-token prediction," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 06-11 April 2025 2025: IEEE, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10890328.

[22] L. Knoedler *et al.*, "Pure Wisdom or Potemkin Villages? A Comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE Step 3 Style Questions: Quantitative Analysis," *JMIR Medical Education,* vol. 10, no. 1, p. e51148, 2024, doi: 10.2196/51148.

[23] I. Didych, "Application of neural network platforms for text-based image generation," 2024.

[24] M. Aydinalp. "Game Design Document for 'Pool Wars' Generated by ChatGPT-4." Zenodo. https://doi.org/10.5281/zenodo.15422946 (accessed 15 May 2025.

[25] L. Giray, "Prompt engineering with ChatGPT: a guide for academic writers," *Annals of biomedical engineering,* vol. 51, no. 12, pp. 2629-2633, 2023, doi: https://doi.org/10.1007/s10439-023-03272-4.

[26] G. Z. Higginbotham and N. S. Matthews, "Prompting and in-context learning: Optimizing prompts for mistral large," 2024, doi: 10.21203/rs.3.rs-4430993/v1.

[27] A. J. Spasić and D. S. Janković, "Using ChatGPT standard prompt engineering techniques in lesson preparation: role, instructions and seed-word prompts," in *2023 58th international scientific conference on information, communication and energy systems and technologies (ICEST)*, 2023: IEEE, pp. 47-50.

[28] P. Denny *et al.*, "Prompt Problems: A new programming exercise for the generative AI era," in *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, 2024, pp. 296-302.

[29] Z. M. Wang *et al.*, "Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models," *arXiv preprint arXiv:2310.00746,* 2023.

[30] Ş. K. Gökçek and D. Akbulut, "Bağımsız Video Oyunlarının Geliştirilme Sürecinde Oyun Tasarımına Yönelik İhtiyaçların, Problemlerin, Benzerliklerin ve Farklılıkların Keşfedilmesi İçin Bir Alan Çalışması," *Sanat ve Tasarım Dergisi,* no. 30, pp. 187-207, 2022, doi: https://doi.org/10.18603/sanatvetasarim.1215230.

[31] B. Dervişoğlu, "Öğrenciler için Dijital Oyun Tasarım Dokümantasyonu Hazırlama Sürecinde Yapay Zeka Kullanımı," *Social Sciences Studies Journal,* vol. 10, no. 12, pp. 2458-2465, 2024.

[32] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, "Likert scale: Explored and explained," *British journal of applied science & technology,* vol. 7, no. 4, pp. 396-403, 2015, doi: 10.9734/BJAST/2015/14975.

[33] OpenAI. "Introducing GPT-4 and its capabilities." OpenAI. https://openai.com/research/gpt-4 (accessed 3 Şubat 2025, 2024).

[34] Google. "ImageFX: AI-powered Text-to-Image Tool." https://imgfx.ai/ (accessed.