ISSN: 2384-8766
https://journal.seriousgamessociety.org/

Article

Can ChatGPT Match Experts? Comparing input for Serious Game Development

Janne Tyni[1], Aatu Turunen[2], Roman Bednarik[1], Juho Kahila[3] and Matti Tedre[1]

[1]School of Computing, University of Eastern Finland, Joensuu, Finland; [2]Department of Environmental and Biological Sciences, University of Eastern Finland, Joensuu, Finland; [3]School of Applied Educational Science and Teacher Education, University of Eastern Finland, Joensuu, Finland
janne.tyni@uef.fi;aatu.turunen@uef.fi;roman.bednarik@uef.fi;juho.kahila@uef.fi;matti.tedre@uef.fi

Keywords:

Game design
Serious games
Game-Based learning
Large Language Models
Artificial Intelligence
ChatGPT

Received: February 2024
Accepted: June 2024
Published: June 2024
DOI: 10.17083/ijsg. v11i2.744

Abstract

This paper investigates the validity of ChatGPT as a tool to generate meaningful input for the serious game design process. Input is collected from game designers, students and teachers via surveys, individual interviews and group discussions inspired by a description of a simple educational drilling game and its context of use. In these mixed methods experiments ChatGPT 3.5 and 4.0 is prompted with the same description to validate findings with expert participants. In addition, the impact on the models' suggestions from integrating the expert's role (e.g., "Answer as if you were a teacher.", "game designer", or a "student") into the prompt is investigated. The findings show that ChatGPT can produce statistically similar input, depending on the group of experts. ChatGPT 3.5 outperforms 4.0 with the student input. The integration of expert's role in prompt is found to be unreliable, and necessary only with game designer input with the version 3.5 of ChatGPT. The practical implications are that multiple ChatGPT versions should be used when collecting input. In addition, it is shown that experts can provide unique insight to the development process. This research opens the discussion on the trustworthiness of ChatGPT generated input for serious game development.

1. Introduction

The advent of easily accessible machine learning tools in the recent years has sparked a pursuit among researchers in the educational field to identify what arduous tasks can be augmented or automated with this new technology. As OpenAI's large language model (LLM), ChatGPT, is a relatively new tool in this field, there is a need for more empirical research [1] to find out its effective uses in education. The use of large language models in the educational sector is a trending topic [2] and if correctly used, the technology could be a prominent catalyst for many reforms [3]. One of the potential applications of effective use for LLM technology is to massively decrease the workload of teachers [1] and other professionals in the field of serious and educational games. It has been suggested that LLMs such as ChatGPT can be powerful tools to design effective learning tasks [1] and to recommend learning content and generate novel learning materials [3]. LLM's

ability to produce content, perform data analysis and to potentially personalize experiences has also been noticed as a potential tool for game development purposes and is anticipated to play an influential role in the gaming industry [4]. In the following sections we will discuss how ChatGPT has been viewed, the current issues in serious and educational game design in addition to how ChatGPT can be leveraged to solve these problems.

1.1 Background

How ChatGPT is perceived by its end users is a topic that has seen increasing attention in the past years. ChatGPT has been used as a part of a collaborative design framework to simulate the human design process and it was received positively by participants [5]. Similarly, A survey study in the field of mental health shows that respondents perceive ChatGPT as a beneficial learning tool [6]. A questionnaire study of university students shows that ChatGPT is viewed in a positive manner [7]. Despite these recent research papers showing that ChatGPT is generally received positively, the current body of literature shows worry over how the LLM's are used and how ChatGPT produced artefacts are received by end users. ChatGPT may not always grasp the intended meaning behind words [1] or it cannot consider the context [1, 4], both of which may result in recommendations that are not appropriate for the skill level of the student [1]. In addition, ChatGPT may not generate novel or imaginative ideas and lacks emotional intelligence [4]. It has been brought to attention by [8] that ChatGPT does not ask clarifying questions; instead, it directly infers what the user wants as an answer.

ChatGPT as a tool has its uses but the literature shows that there are issues that could impact its effectiveness in the design of serious and educational games. It is not recommended to exclusively rely on the power of ChatGPT [8] but instead evaluate the advantages and disadvantages to decide whether or not to use ChatGPT for game projects [4]. Considering that the design of tutoring activities is not a simple task and components like the game experience as a whole and the cognitive load of the learner has to be taken into account [9], these issues could potentially influence the application of LLM's to generate useful input (suggestions or feedback) for educational game development. The use of LLM's to aid in the design of input for educational games would be helpful, as the task to produce an engaging end product is a difficult one [10]. There is little empirical data available on how to design games that effectively meet the educational intentions they were created for [11].

The core principles of educational game design are one proposed way to design educational games that meet its educational intentions. "Game motivation" and "game thinking" are coined as some of the core values of educational games [9]. These are defined as "the most basic and most operational value, emphasizing the application of games to learning to stimulate students' learning motivation and make the learning process more attractive and interesting" and the ability to apply elements from games to teaching, respectively. Similarly [12] attributes fun, motivation and engagement as key aspects of producing engaging educational games.

One way to add the "fun factor" to serious and educational games is gamification. Gamification has been defined as a way to encourage expected behaviors via the application of elements or mechanisms from game design in a non-game context [13]. Gamification can be used to introduce the "game motivation" and "game thinking" as introduced by [9] and to increase the engagement of the students. Studies support the view that tasks with game elements are viewed as more attractive than their non gamified counterparts [14, 17]. However, the body of literature in the use of game elements in learning has shown inconsistent results [13,14]. A recent research paper publication suggests that this inconsistency in gamification results may be because of the gamification elements such as badges, points and achievements are added without any consideration as to why and for what purpose [15].

Research in the field of rewards in educational and serious games has found that gamification, game design and game-based learning are prevalent topics in the body of knowledge in this field of research [16]. A bibliometric analysis focusing on gamification trends in higher education between the years 2013 and 2022 shows that the number of studies in the field is steadily growing [18].

In short, ChatGPT as a tool is viewed positively. However, the credibility of ChatGPT generated artefacts is not yet defined. Studies suggest that ChatGPT should not be used haphazardly but with

caution. One difficult task where ChatGPT could provide useful is the design of serious and educational games. The possibility to produce motivating and educational games has been researched by several studies, but no single solution has yet been discovered. The addition of "fun factor" elements such as gamification to educational games has been suggested as one potential solution. The quest for the true potential of educational games is still ongoing [19], and the field of educational game design is an active field of research [10].

1.2 Knowledge gap

Game design research suggests that the development of a successful educational game involves having stakeholders participate in the design process and having the end product be simple [12]. Instead of basing the design of an educational game on only theory, collaboration with educators and other professionals in the game industry in addition to students might provide better results [19]. A study published in 2023 suggests a participatory methodology to produce "engaging, effective and learning-friendly games" [20]. In this methodology, a multidisciplinary team consisting of software developers, pedagogical experts, game designers and the end-user is suggested. This team will then proceed to interact and provide feedback from all disciplines to produce a game. The process of identifying the optimal methods to keep the player motivated to play the game is "very long and requires different expertise from all the stakeholders" [12].

There are multiple ways to gather input from stakeholders and the types of input may vary depending on the serious or educational game that is being developed. One way to gather input is to hold focus group interviews. Interviewing participants for ideas can yield input suggestions such as avatars, in-game currency and the possibility to trade [21]. Other games have implemented features like experience points, levels and boosters [22]. The input can also be vague, such as co-operation, puzzles [12], feedback or an end goal to a game.

Feedback for an educational game in addition to an end goal are proposed as engaging gamification elements [13] and defined as educational game design principles in [10]. A possible end goal for a game could be achieved through a story, similar to an educational story driven adventure game presented in [23].

As discussed, the input suggestions from experts can range from entire game modes to simple in game rewards. It has been noted that in mobile games the games made for recreational purposes have more variety in reward types when compared to games made for education [24]. There are plenty of studies that have explored different methodologies to collect input for the development of serious and educational games. However, this area of interest has not yet much research that tries to leverage the use of LLM's such as ChatGPT.

A study [25] tested prompting ChatGPT 3.5 as a tool to brainstorm ideas for board games. The study found that ChatGPT can be used as a valuable tool to design and further develop board game ideas. In addition to board game ideas ChatGPT has also been harnessed to produce ideas for educational escape room design [26]. In this study, the prompt was engineered to "Act as a …" persona or a character. In this case, ChatGPT was asked to act as an educational game room designer who used a specific framework. If an LLM, such as ChatGPT can be shown to produce input that is similar to groups of experts then the workload of serious and educational game designers and developers can be dramatically lessened.

To summarize, involving stakeholders in the design process of serious and educational games is important. This way, the end product will more likely meet its intended purpose. There are multiple ways to gather input from stakeholders, and the input can range from small concrete changes to larger ideas that are less easily defined. The body of literature where ChatGPT is used to aid in the process of collecting input from experts is still in its infancy. The existing studies that leverage ChatGPT in serious and educational game design by no means answer to all of the questions that are left unanswered; such as the credibility of ChatGPT generated input and the usefulness of a stakeholder (or expert) persona in a prompt.

1.3 Objectives of the study

In this study the potential of ChatGPT for generating input for educational game development purposes is investigated. The two available versions of ChatGPT (4.0 and 3.5) are chosen for this

study, as the popularity of ChatGPT makes it well known in the field. In addition, ChatGPT prompt research in game design ideation is limited; gaps exist in studies such as [25] which included ChatGPT 3.5 but not 4.0 and [26] in which a "Act as a…" persona or character was used but the results are not compared to human experts. The number of studies that investigate ChatGPT's credibility as a source of expert input is limited. This leaves open questions regarding the differences between the two versions of ChatGPT and the credibility of the input.

As the field of prompting ChatGPT for educational game development input as a persona is limited, the study follows similar methodology to the conventional content analysis [27] and the inductive content analysis process [28]. The results are compared with those obtained from interviews with experts, considering both scenarios: with and without an expert prompt provided to the ChatGPT. The results are further examined to answer the following research questions:

(1) How does the input from data sets collected from ChatGPT differ from input collected from experts?
(2) How does the input from data sets collected from ChatGPT differ from each other when given the instruction to answer as a persona?
(3) Can the input collected from ChatGPT be utilized to enhance the future design and development process of educational games?

By comparing the suggestions given by expert participants with the data collected from ChatGPT this study advances the field of educational game design by exploring the possibility of using ChatGPT as a tool to reduce workload and to provide more efficient and relevant input for educational game development.

2. Methodology

Keeping the game simple leads to successful adoption results [12]. Touch-type educational games significantly improve learning motivation [29]. To accommodate these principles, the present study used a touch-based mobile educational game based on bird recognition and memorization. Similarly, [17] suggests that gamified mobile learning technology increases learning engagement.

The mechanics of the game are as follows: the game shows a picture of a bird, which is then instructed to be moved to the correct name of the bird. There are three options to choose from. After moving the bird to a name option, picture of another bird or a different picture of the same bird species appears on the screen. This task is to be completed before a timer, which is presented on the same screen, runs out and ends the game. After the game ends, a window showing the correct name of the last incorrectly guessed bird picture is shown. The game's objective is to name all 20 individual birds three times in a row.

In this research, a primary prompt was produced by the researchers based on the game and the context of use. was explained to the ChatGPT models. The input provided by the ChatGPT was collected and analyzed. The same prompt was given to experts in the form of a survey and the game was demonstrated to some of the experts and they were interviewed for input. Adhering to the conventional content analysis methods [27], the surveys and interviews were held with open-ended questions. The data was gathered and analyzed similarly to the principles of inductive content analysis [28]. Finally, the results were compared to the input collected from the ChatGPT and expert participants.

2.1 ChatGPT data collection

A simple verbal explanation was formed based of the functionality from the simple educational game. The prompt that was used for the ChatGPT is as follows:

```
You are shown a demo version of an educational video game. The game
screen shows a picture of a bird in the middle of the screen. There are
three text boxes above the bird. Each textbox corresponds to a bird
name. The same name can appear multiple times. To the left of the
screen, there is a bar that is diminishing during gameplay. This bar
```

```
symbolizes a timer. When the timer runs out, the game ends. At the end
of a game, the game presents a small window in which the picture of the
last incorrectly guessed bird and the correct name is shown.


The objective of the game is to move the bird picture to the correct
name. When a bird is moved to the correct name, a symbol that says
"Correct!" is shown for a second. Afterward, a random picture of another
bird species appears on the screen. There's a set amount of different
bird species (with different images for each bird species). The picture
of a bird is removed from the pool of possible birds once it has been
correctly dragged to the correct name three times in a row. Once all 20
birds have been named correctly, the game ends.


The game is meant for biology students studying the names of birds.
There are a total of 200 birds that the students need to study for an
exam.


What kind of changes or additions would you make to the game?
```

Data was collected with the prompt as is to ChatGPT 3.5 and ChatGPT 4.0 to collect a "baseline" of answers. The text "Answer as if you were a teacher.", "Answer as if you were a game designer." or "Answer as if you were a student." was added to the end of the prompt to collect the data sets with the different personas. To make sure that the previous prompt would not affect the input given by the LLM a new thread (or tab) was opened in order to collect data with the variant prompt. The verbal explanation was then delivered to the different ChatGPT versions (3.5 and 4.0) via the website https://chat.openai.com/. The OpenAI service offers a button to "Regenerate response", which was used to regenerate the responses 50 times for both collected data sets in order to saturate the pool of possible answers. The ChatGPT client from the website always responded with a numerated list of suggestions. The responses from the two versions were collected between the dates 05.06.2023 and 11.06.2023.

The data sets were sorted in a way that the word combinations that meant the same thing, such as "Bird Calls", "Bird Calls and Sounds", "Bird calls or Songs", were categorized under the same category "Bird calls or sounds". The items from the responses were listed for further refinement. A total of 72 unique suggestion categories were then gathered. The data sets that held multiple suggestions separated by the "and" or the "or" conjunction were separated for categorization. Whether or not the suggestion categories were present in each of data sets was tested by using the chi-square test (X2) in R-software (version 4.1.3) and visualized using corrplot-package [30].

2.2 Expert input collection

A survey with the same prompt that was given to the ChatGPT was delivered to various internet forums, contacts and study group chats to gather input from experts from the 31st of August to the 29th of September 2023. The first question in the survey was "Are you a" followed by four options: "Student", "Teacher", "Game designer" and "Something else, what?:". The last option allowed the participants to describe themselves in an open answer box. After that the prompt was given followed by an open answer box. These groups of experts were chosen because they are key stakeholders in the design process of an educational game (as the end users, educators and designers of the product).

In addition, six participants who identified as being of the teaching profession and a game designer were interviewed for this study. The interviews were conducted during 1 hour sessions. The session began by the researcher presenting the educational game and explaining its functions in a similar way to the prompt used for ChatGPT data sets. The researcher continued by asking questions such as "What kind of changes or additions would you make to the game?" or "What did

you like or dislike about the game?" and follow-up questions such as "Why, can you elaborate?". Additional leading questions regarding the motivation of the player to play the game based on whether or not they were a biology student studying for a test or a hobbyist trying to learn the names for fun were provided by the researcher if they were not suggested by the participants. This was performed to evaluate whether the responses would vary depending on what kind of user group the educational game was for. The sessions were recorded and transcribed. Notes about the thoughts and suggestions of the participants were listed and quoted.

3. Results

The first subsection presents the data collected from both the ChatGPT versions (4.0 and 3.5) with and without a persona prompt. The result of the chi-squared test is presented. The subsection that follows presents the analyzed thoughts and quotes from the expert interviews. The suggestions gathered from interviews and the survey are listed and presented. The last subsection compares the results from the first two subsections in terms of presence and the number of suggestion categories.

The categories were formed by the researchers by grouping the individual input and codes into larger categories. Table 1 shows the four categories which were formed and the subcategories that further explain the contents of the categories. Examples are provided from the individual suggestions to show what kind of suggestions were included in each of the categories. The data analysis was done with these suggestion categories.

Table 1. Categories and subcategories that were collected from the ChatGPT and expert data sets.

Category	Subcategory	Example
Game modes and mechanics	Game modes, Game mechanics	Multiplayer, Virtual field trip Day and night cycle, Ability to retry missed birds
Learning and educational content	Learning content, Classification and categorization, Integrations, Learning features, Analytics	Bird calls or sounds, Scientific names Classification, Categorization Integration with classroom activities Adaptive learning features, Feedback Game statistics and analytics, Progress tracking, Categorization
Accessibility and user interface	Customization features, Accessibility, Aesthetics	Bookmarking feature Multilingual support, Platform compatibility Scenic backgrounds
Gamification		Achievements, Scoring system

3.1 Input from human experts

Input from the experts was collected via a survey (N=88). In addition, some of the participants across all groups were interviewed in group and individual interviews. The expert interviews began with the researchers giving the participants a verbal description of the game (as presented in 2.1). After hearing the verbal description, the participants tried out a mock up of the game (Figure 1) until they felt that they had understood the concept and the flow of the game. The first suggestions often were produced within minutes after testing the game. The first concerns of the participants were often accessibility options such as the ability to zoom the picture of the bird and different methods to offer feedback to the students after choosing the bird name incorrectly.

Figure 1. A mock up based on the description of the educational game. At the center, there's a picture of a bird. There are three name options to drag the picture of the bird, and a timer on the left side of the screen that is slowly depleting.

"[...] incorporate a feedback mechanic when the player makes too many mistakes there could be like a pop up window coming up like "the bird's name is this"".
—Teacher participant

When asked what the participants liked or disliked about the game the impressions in general were positive. The timer feature in the game was often discussed as a method to keep the player from either cheating or a way to keep the player (or student) from looking up the bird pictures online or from other sources to promote learning. Similarly, a game designer participant agreed that there should be some kind of control over the possibility that a person could pass a level on pure luck alone. The student feedback for the timer was mixed. Some did not raise any issue with it, but other students stated that the "Timer starts to build up a lot of pressure on the student" and hoped that the timer could be an optional feature in the game.

"The timer locks the player to the level so that the player can't google the name of the bird. If we want the player to learn it should be within the system, so I think the timer is a good idea for that. You don't want to depend on external sources to initiate the learning. So that's why the timer is a good idea".
—Teacher participant

"Loopholes are definitely good. Players that can find loopholes from games are worth more than gold, especially in game design. However, loopholes in educational games are not as good (as they can incentivize not learning but skipping the content instead)".
—Game designer participant

After the researcher asked about the specific need for rewards such as badges, achievements or monetary rewards. The teachers were stunned for answers. However, after asking whether or not the game needed rewards if it was designed for hobbyists instead of students the teachers often were more willing to add different "fun factor" and reward mechanics. When inquired further about the reasoning behind this the teacher participants often mentioned the intrinsic motivation of the student user group.

"A biology student shouldn't need any specific rewards. It's a part of the studies so the study credits should be reward enough".
—Teacher participant

The game designers and students were more eager to add different kinds of gamification and reward mechanics into the game. A game designer thought that by adding different features (such

as giving points and the possibility for “streaks” when multiple correct guesses were made in a row) and changing the “core game loop” of the game could help players or students stay interested in the game for longer periods of time. Similarly, student participants felt that adding a scoring system might help the game feel more exciting. “By adding a scoring system u can make the game feel more competitive”, a student stated in the survey.

“Change to a score based system. […] The bird should be kept in the pool of answers and should grant some amount of points. The faster player drags a bird to the name the more points they get, which promotes remembering the birds using quick thinking and reflexes”.
—Student participant

When asked for thoughts about using ChatGPT as a tool for improving educational games the responses were positive. ChatGPT was seen as a way to gather suggestions for improvement. When discussing the usage of ChatGPT in serious games feedback collection it came apparent that it was not seen as a tool that should be used only by itself. ChatGPT was seen as an additional tool to the educational game development, not as a replacement for more traditional feedback gathering methods such as presenting the game to actual human beings, experts and end users. Doubts were expressed about the validity of the input given by ChatGPT, as it was inferred that it may not be able to grasp the whole context of the educational game.

“I think ChatGPT will be valuable. – I think it will give you a sort of average response among education professionals about what you’re asking”.
—Teacher participant

“I think it could give you some ideas and suggestions that are valuable. But maybe you could see that
they are not very fitting to what you are trying to do”.
—Teacher participant

The suggestions that were provided by the expert participants often came evenly throughout the interview sessions. In addition to the interviews, input for the development of the educational game were collected via a survey. Combined, the input from the interviews and surveys produced 148 individual suggestions. The suggestions were categorized by the researchers. Figure 2 shows the most frequent categories of input for each of the expert data sets in descending order.

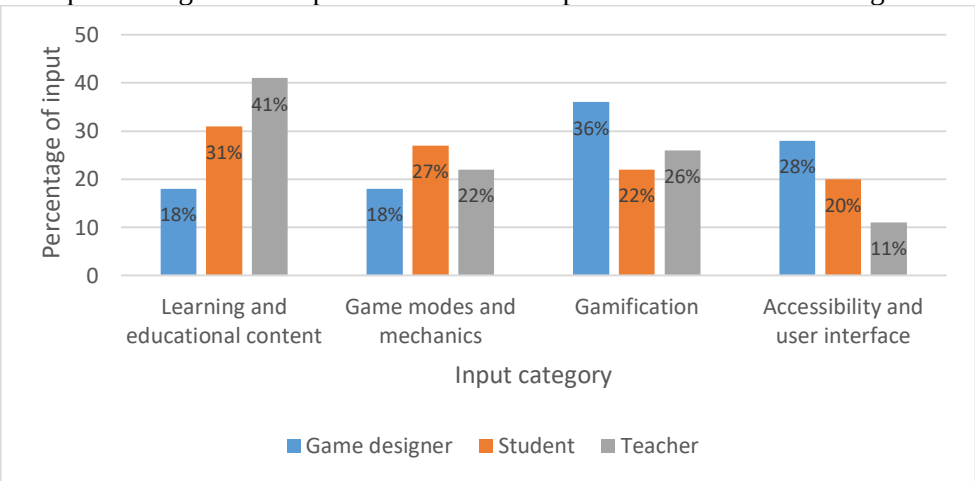

Figure 2. This Figure presents the percentages of the categorized input suggestions produced by the experts.

As can be seen from Figure 2 the student and teacher data sets are similar in order of numerical frequency with the student data set having more input in the "Game modes and mechanics" category than in "Gamification" and vice versa for the Teacher data set. The game designer input had the

category "Gamification" represented with 36% of the total input in that data set, when in the other two data sets it held only 22% and 26% of the input.

3.2 Input from ChatGPT

After prompting each of the ChatGPT versions with the prompts and regenerating the answer for 50 times a total of 1455 individual suggestions were produced. As can be seen from Table 2, ChatGPT version 4.0 produced more suggestions than its 3.5 counterpart, excluding the prompt to answer as a game designer. The most populated input categories by all the ChatGPT 3.5 (excluding ChatGPT 3.5 game designer) data sets were, in descending order: "Learning and educational content", "Game modes and mechanics", "Gamification" and "Accessibility and user interface". The ChatGPT 3.5 Game designer data set appears to have more input suggestions in the category "Game modes and mechanics" when compared to the "Learning and educational content" category, which was the numerically most populated category in the other ChatGPT 3.5 data sets. Figure 3 shows that the first three categories in descending order held 21% to 38% of suggestions in individual categories while the last category, "Accessibility and user interface", had suggestions from 10% to 15%.

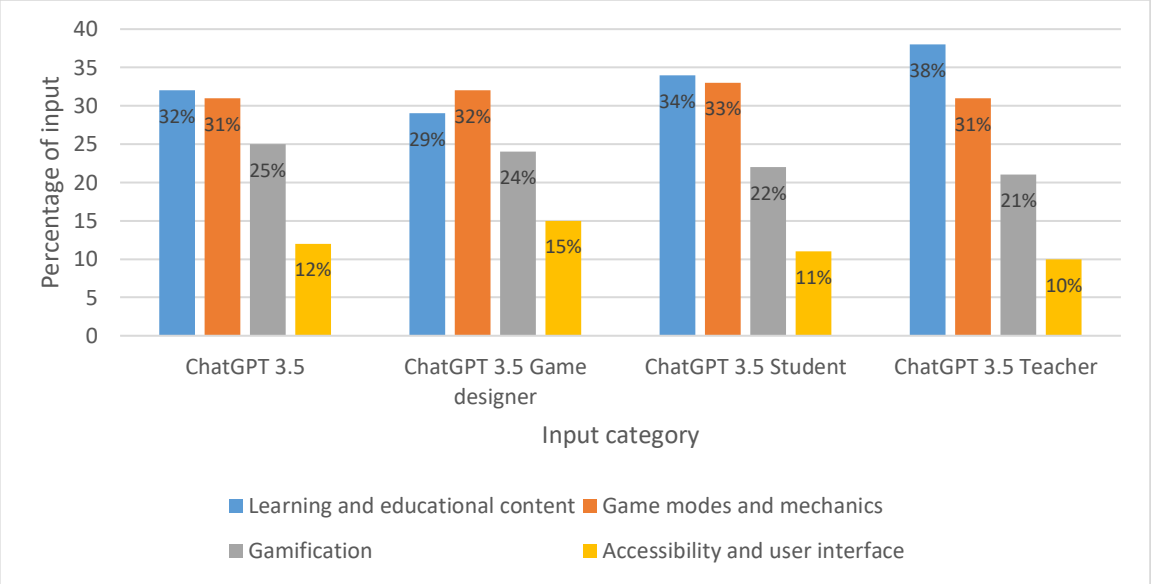

Figure 3. This Figure presents the percentage of the categorized input suggestions produced by ChatGPT 3.5.

The most numerically populated input categories in the ChatGPT 4.0 data sets in descending order were the same in all the suggestion categories; "Learning and educational content", "Game modes and mechanics" and "Gamification" followed by "Accessibility and user interface". As can be seen from Figure 4, the most populated category, "Learning and educational content", has 37% to 51% of the suggestions in each data set collected from ChatGPT 4.0. The percentage of suggestions in the category "Game modes and mechanics" range from 26% to 29% in each of the ChatGPT 4.0 data sets. The last category in these data sets holds 5% of all categorized suggestions in all data ChatGPT 4.0 sets.

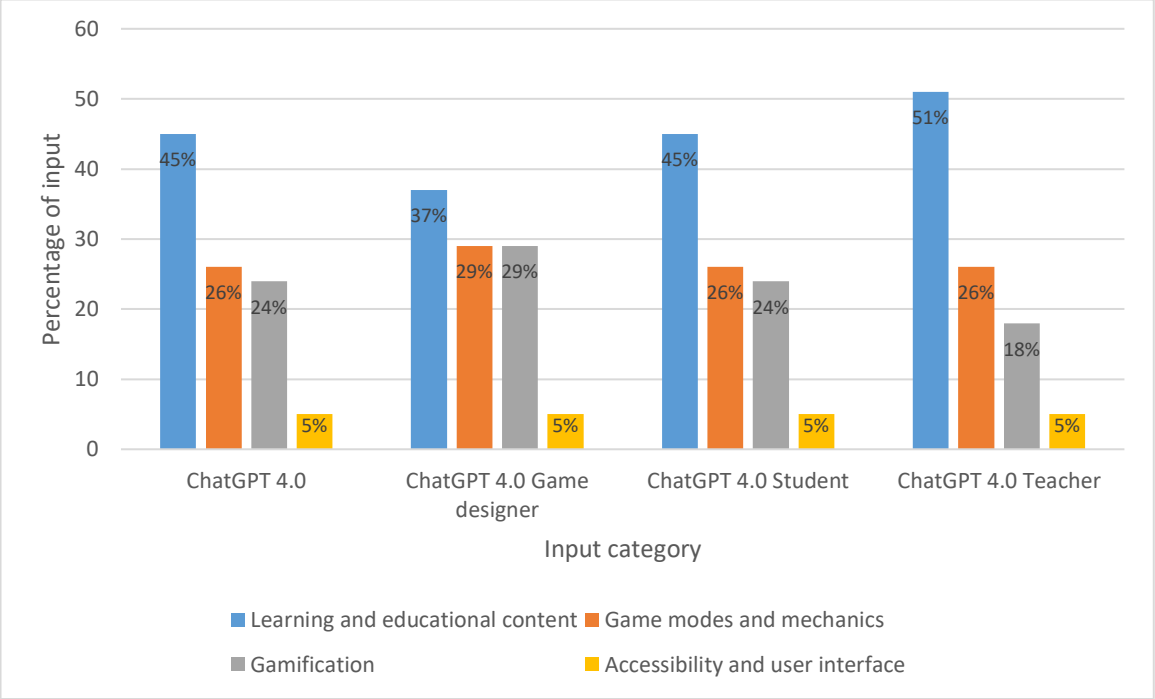

Figure 4. This Figure presents the percentages of the categorized input suggestions produced by the ChatGPT 4.0.

Table 2. The number of individual suggestions produced by each ChatGPT version and prompt. The numbers are in descending order.

Prompt	Number of suggestions
ChatGPT 4.0	275
ChatGPT 4.0 Student	202
ChatGPT 4.0 Teacher	197
ChatGPT 3.5 Game designer	174
ChatGPT 4.0 Game designer	160
ChatGPT 3.5 Teacher	157
ChatGPT 3.5	145
ChatGPT 3.5 Student	145

3.3 Comparison between ChatGPT and human experts

After removing duplicates, but without categorization, all of the ChatGPT data sets combined offered 1455 individual suggestions. Expert input yielded 148 individual suggestions for the improvement of the game. After categorizing the ChatGPT suggestions and comparing them to the data from experts (Figure 5) it can be seen that the ChatGPT produced many similar responses when compared to the input from the expert participants. Chi-square test was conducted on all the data sets with 200 replicates used in the Monte Carlo test ($\chi 2$ = 182.47, df = 3, p-value = <0.001). Figure 5 shows positive and negative associations between the data sets and suggestion categories. This implies that there is statistically significant difference in the number of suggestions when comparing the data sets with different prompts. From this figure it's possible to deduce that all the

ChatGPT 4.0 data sets are negatively associated with the "Accessibility and user interface" category and all other data sets are positively associated with that category. The ChatGPT 3.5 data sets are negatively associated with the category "Learning and educational content" when all its ChatGPT 4.0 counterparts, except the ChatGPT 4.0 Game designer, are positively associated with the same category. Similar division can be seen in the "Game modes and mechanics" category, where ChatGPT 3.5 data sets are positively and ChatGPT 4.0 data sets are negatively associated, except for ChatGPT 4.0 Game designer data set. All the expert participant data set categories show negative association with this category.

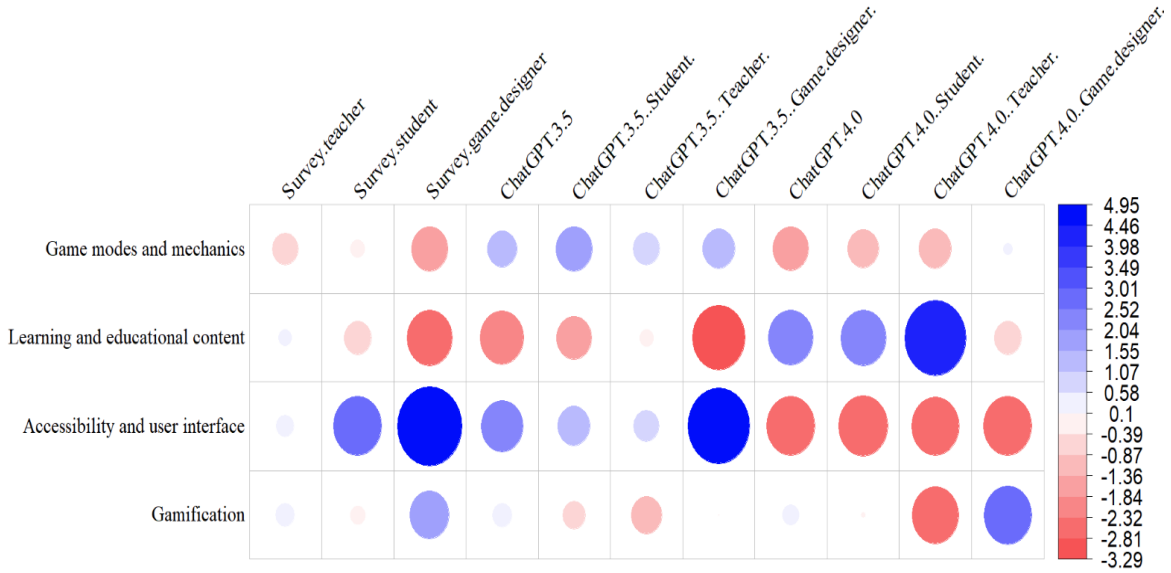

Figure 5. Chi-square test comparing the answers of both ChatGPT 3.5 and 4.0 with the expert participants. Positive association is specified with the blue color. Negative association is specified by the red color. Strength of the association is represented by the size of the dot.

The comparisons of the pairs of data sets were conducted with the Chi-square test and the results can be seen in Table 3. The tests were done with the variables provided in Table 4. Statistically significant comparisons ($p < 0.05$) suggest that there is difference between the abundance in answers between the prompts. Table 3 shows that the data set gathered from students was statistically different from the data sets collected from ChatGPT 4.0. The same is true when comparing the data set collected from game designers with the data sets from all the ChatGPT 4.0 data sets except the one without prompt.

Table 3. Paired comparisons of Chi-square test between the paired data sets. G is shorthand for game designers, T is for teachers, S is for students. 3.5 For ChatGPT 3.5 and 4.0 for ChatGPT 4.0. If the value is below <0.05 the data sets have statistically significant difference in content when compared to its pair. P-values are corrected for multiple testing by us

Data set	4.0 G	4.0 T	4.0 S	4.0	3.5 G	3.5 T	3.5 S	3.5	G	S
T	1	1	1	1	1	1	1	1	1	1
S	0.0282*	0.007*	0.013***	0.0217*	1	1	1	1	1	
G	<0.001***	<0.001***	<0.001***	1	0.111	<0.001***	0.001**	0.010*		
3.5	0.054	<0.001***	<0.001***	<0.001***	1	1	1			
3.5 S	0.080	<0.001***	0.005**	0.004**	1	1				
3.5 T	0.119	0.008**	0.164	0.143	0.801					

3.5 G	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***
4.0	1	1	1		
4.0 S	1	1			
4.0 T	<0.001***				

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

Table 4. Paired comparisons of Chi-square test between the paired data sets. G is shorthand for game designers, T is for teachers, S is for students. 3.5 For ChatGPT 3.5 and 4.0 for ChatGPT 4.0. The Chi-square tests in Table 6 were done with the values provided in this table.

Data set	4.0 G	4.0 T	4.0 S	4.0	3.5 G	3.5 T	3.5 S	3.5	G	S
T	2.87	4.47	2.69	2.22	3.15	1.70	2.16	1.95	8.22	2.27
S	17.67	20.71	19.34	18.23	1.38	5.44	4.43	3.28	5.73	
G	44.59	57.03	51.26	48.75	14.78	27.95	24.59	19.85		
3.5	16.30	41.18	26.93	26.55	2.66	4.17	1.27			
3.5 S	15.48	31.17	21.29	21.60	5.35	1.35				
3.5 T	14.63	20.26	13.94	14.23	10.53					
3.5 G	31.13	58.89	44.74	44.58						
4.0	9.12	6.98	0.19							
4.0 S	9.45	5.68								
4.0 T	27.76									

Table 3 shows that the ChatGPT 4.0 (Game designer) data set was statistically different from the expert game designer category in addition to the ChatGPT 3.5 (Game designer) and ChatGPT 4.0 (Teacher) categories. The ChatGPT 4.0 (teacher) categories are statistically different from the students, game designers and all the ChatGPT 3.5 data set categories. The ChatGPT 4.0 student category is statistically different with the students, game designers and all the ChatGPT 3.5 categories.

4. Discussion

4.1 Main Results

In this study, input from expert participants and ChatGPT was collected and analyzed to investigate the trustworthiness of ChatGPT input for the design process of serious and educational games. The input collected from both expert participants and ChatGPT was categorized by the researchers into four different main categories: “Game modes and mechanics”, “Learning and educational content”, “Accessibility and user interface” and “Gamification”. The numerical frequencies of these input categories from expert participants are shown in Figure 2. From this table it can be seen that the

numerical distribution of input from students and teachers share the following commonalities; the most common input is in the category "Learning and educational content" and the least common input category is "Accessibility and user interface". Where these two expert groups differ is that the teacher participants have 4% more input in the "Gamification" category when compared to the "Game modes and mechanics" category. The opposite is true for the student participants, who provided 5% more input in the latter category than the former. These two expert categories are heavily tied to the concept of learning, so it comes as no surprise that the most frequently given input was related to education.

Game designers provided input that is distributed as follows, in descending order: "Gamification", "Accessibility and user interface", "Learning and educational content" and lastly "Game modes and mechanics". This data shows that there are similarities between the input provided by students and teachers, but this is not the case with input from game designers. The prevalence of input in the "Gamification" and "Accessibility and user interface" categories may very well be from the fact that this expert group is the most familiar with concepts and practical applications of these two concepts. The game designer participants were also the expert group most concerned with the user experience and the end users having fun while engaging with the game.

The frequency and most favored frequent input categories of ChatGPT with the different versions and prompts can be seen from Figures 3 and 4. When looking at the input categories of ChatGPT 3.5 (Figure 3) it can be seen that the distribution of input is more homogenous when compared to expert participants (Figure 2). Except for ChatGPT 3.5 with the game designer prompt, all the other data sets show that the input is distributed as follows, in descending order: "Learning and educational content", "Game modes and mechanics", "Gamification" and "Accessibility and user interface". The ChatGPT 3.5 with the game designer prompt provided slightly (3%) more input in the "Game modes and mechanics" input category than "Learning and educational content" category. Otherwise, the distribution of input is similar in frequency between all the data sets collected from ChatGPT 3.5. All the data sets collected from ChatGPT 4.0 followed the same order of frequency in input categories as the majority of ChatGPT 3.5 data sets.

When comparing the frequency of input between ChatGPT 4.0 (Figure 4), ChatGPT 3.5 (Figure 3) and the expert participants (Figure 2) it can be seen that the input between the teacher and student participants share the same numerical frequency as do most of the input prompted from ChatGPT 4.0 and ChatGPT 3.5 with or without the persona prompt. It is interesting to note that this holds true even though ChatGPT 4.0 in general gives more individual suggestions when compared to ChatGPT 3.5 (Table 2).

In terms of numerical frequency of input categories, this study provides evidence that ChatGPT can be trusted to produce input similar to expert participants (research question 1). However, similarity in the numerical frequency of input categories themselves is not enough to say whether ChatGPT input can be trusted when compared to input from real experts. To accommodate this need, paired comparisons of Chi-square test were calculated with all the data sets.

Table 3 and Figure 5 show several key facts that help assess the differences in input from ChatGPT and experts. First, the teacher data set is not statistically significantly different from any other data set. This finding has direct ramifications to the design and development of serious and educational games. In practice this shows that participant groups, even if they are from different areas of expertise, can provide input that is statistically similar to experts from other domains. This also suggests that with some expert groups and prompts the addition of a persona prompt or using a different version of ChatGPT does not produce statistically different input. This finding should be taken into account especially if statistical analysis tools are used.

The second key finding of this study is that when comparing input from students with all the ChatGPT prompts, it can be seen that ChatGPT 3.5 does produce input that is not statistically significantly different (Table 3). However, all of the ChatGPT 4.0 prompts have statistically significant differences in comparison. This finding has implications to direct the practical use of ChatGPT in input collection for serious and educational games. In practice, this result shows that it is possible that the version of ChatGPT matters when prompting for input. Therefore, it is recommended to prompt different versions of ChatGPT, especially when statistical analysis tools such as a paired Chi-square test is used.

Third, the "answer as if you were a game designer" prompt produces statistically noticeably similar input when compared to other prompts with the ChatGPT version 3.5. However, ChatGPT 4.0, with

any of the persona prompts, does not show similarity to the expert game designer input. ChatGPT 4.0 without a persona prompt, however, is statistically similar. This finding shows that with the ChatGPT version 3.5 the persona prompt "answer as if you were a game designer" should be used when prompting for input. However, when using ChatGPT 4.0, a persona prompt should not be used at all. The "answer as if you were a game designer" prompt was the only persona prompt that succeeded in meeting its expectations (producing input similar to game designers). This was only the case with ChatGPT version 3.5.

Taking all these findings into account we have three very different results. First, the teacher data set is statistically similar to all of the other data sets. Second, the input collected from students is similar only to input produced by ChatGPT 3.5. Third, the input from game designers is only statistically significantly similar to ChatGPT 3.5 with the corresponding "answer as if you were a game designer" prompt and the ChatGPT 4.0 without a persona prompt. These findings show that the persona prompting method, as it was presented in this study, worked with only a specific group of experts (game designer) and with a specific version of ChatGPT.

Thus, this paper provides evidence that the persona prompting method is not reliable for all expert groups but can work with very specific prompt, ChatGPT version and expert group (research question 2). The results also show that when using statistical analysis tools to compare prompts, the use of different ChatGPT versions is recommended, as the results can vary depending on the version. It is important to note that this study also shows that the newer version of ChatGPT (4.0) does not always outperform the previous version (3.5).

4.2 Quality and relevance of ChatGPT input

ChatGPT produced a variety of input and it was distributed in a similar manner within ChatGPT versions (Figures 3 and 4). The most frequent input from the ChatGPT data sets combined were, in descending order: learning resources, progress tracking, difficulty levels, multiplayer, feedback, customization options, hints and clues, learning mode, bird calls or sounds and mini-games and quizzes. The least frequent input were, in ascending order: multimedia integration, images of birds in different poses or stages of life, integration with study materials, daily or weekly challenges, virtual field trip, narrative element, offline access, expansion packs, game statistics and scientific names.

The expert participants combined show a tendency to produce input in several key areas. The expert participants combined heavily favored the possibility of multiple game modes, accessibility features, a scoring system, hints and clues, adaptive learning features, game statistics, images of birds in different poses or stages of life, leaderboards and learning resources.

Some of the most frequently found input from ChatGPT overlaps with the input collected from expert participants, but there are key differences. Both ChatGPT and experts produced input for the addition of learning resources, hints and clues and alternative ways to play the game. The expert participants were more likely to ask for different ways to play the game directly, where ChatGPT produced direct suggestions such as "quizzes and mini-games". The input that highlighted the need for learning resources was clear from all data sets. The inputs share these similarities, but there are notable differences.

ChatGPT and the experts had some differences in the variety of input. ChatGPT produced input that favored progress tracking, difficulty levels, multiplayer, customization options, hints and clues, learning mode and bird calls or sounds more than experts. Experts favored leaderboards, points, additional modes to play the game, adaptive learning features, and more pictures of birds in different poses or stages of life more than ChatGPT did.

The comparison of input collected from ChatGPT and Experts reveals a variety of insights. First, both ChatGPT and experts show worry over the learning materials provided in the game concept. ChatGPT heavily produced input for the addition of learning materials (Figures 3 and 4). The input from experts show worry over the students having access to the learning materials before engaging in a play session.

Second, both ChatGPT and experts produced input that aids the learning activity itself: hints and clues. All the data sets show that the game should nudge the player towards the correct answer if the player does not know the correct answer straight away. A notable difference to this is that comparatively experts had more interest in adaptive learning features specifically (adapting to the players skill level) when compared to ChatGPT, which produced input that suggested adding hints and clues in general.

Third, both ChatGPT and experts produced input that suggested that there should be alternative ways to play the game. ChatGPT produced individual suggestions, where expert participants just wished for more variety. This input from experts shows that a typical "matching game" where a picture is connected to a name, can be seen as a idea that has been presented enough, and new innovations should be presented. ChatGPT did not produce input that suggested that the genre of the game was boring, but it did suggest that there should be options to play the game in different ways.

Fourth, the differences in the ChatGPT and expert data sets show that the experts produce more input that is directly related accessibility features and gamification ("fun factor" elements such as points, leaderboards, and additional modes to play the game). This is possibly due to the prevalence of gamification elements in educational games.

In conclusion, it can be said that most of the individual suggestions that were collected from ChatGPT are qualitatively similar to expert participants input. Both ChatGPT and the experts produced actionable input such as learning resources, hints and clues and suggestions for additional ways to play the game. Additionally, both ChatGPT and expert participants occasionally produced input that suggested something unique (and perhaps a bit more demanding to implement) such as a completely new game concept. For example, a bird zoo (idea from experts) or a virtual field trip (input from ChatGPT). The input from ChatGPT is highly relevant to the prompt and is of similar quality to expert participants. In addition, it is worth noting that expert participants do seem to have more of a focus in accessibility and gamification elements when compared to ChatGPT.

4.3 Limitations

The study comes with several limitations. As the data from ChatGPT was collected between the dates 05.06.2023 and 11.06.2023, the inner workings of OpenAI's LLM's (large language models) are subject to change when time passes on. The prompt that is used for gathering input from LLM's greatly influences the given output, so using a vastly different prompt might yield far better or far worse results.

As the resources for this study were limited and the type of persona prompt and query are unique, this study is limited only to ChatGPT. Despite the decent number of experts participating (n = 88), expert interviews that were held had a relatively few participants (n = 7) and the opinions of experts may vary when different participants from various backgrounds are taken into account. The experts that participated in this study were self reported, so a study that has certain criteria to define a specific expert in a field (for example, minimum 10 years of work experience) may produce different results.

Even though the data sets were collected from the ChatGPT versions by regenerating the answer by 50 times, it may be possible to see different distribution of input categorizations when prompted differently or regenerated by more or less than fifty times. The categorizations made for the study by the researchers may be different for researchers who view the suggestions from the data set in a different way.

4.4 Comparison with previous research

The input from the ChatGPT had common themes that emerged when compared with the input obtained from expert interviews. Input for adaptive learning features, customization options,

difficulty levels, feedback, hints and clues among other features that cater to different learners were found not only in ChatGPT data sets but the expert interviews as well. LLM's such as ChatGPT as a tool to improve the game design process was viewed positively in the expert interviews, similar to the response to the tool used in [5], questionnaire by [7] and a survey by [6]. There were concerns that the tool would not be able to correctly grasp the whole context of the educational game and give suggestions that are too general in nature to be used in specific educational games. This was a thought similar to what has been previously suggested, which was that the model may not generate truly novel or imaginative ideas [4]. As can be seen from the Figure 2 the data shows that when using both the ChatGPT 3.5 and ChatGPT 4.0 the same prompt, when compared to expert participants, can produce different input suggestions.

The findings of this study provide valuable insight for the usage of ChatGPT versions 3.5 and 4.0 for the purpose of using collecting input for the development of an educational video game. As can be seen from Table 6 the ChatGPT 3.5 and ChatGPT 4.0 data sets have commonalities and differences when compared to expert participants (with the exclusion of the data set collected from teachers which appears to be statistically similar to all other data sets). The similarities that occur within the ChatGPT versions may be because of internal differences between the ChatGPT versions 3.5 and 4.0, but this does not completely explain why the data sets with some of the persona prompts differ from each other. The results from the data sets seem to suggest that simply adding a persona prompt to answer as if the ChatGPT was enough to produce the sensitivity to changes in prompt phrasing or multiple attempts for ChatGPT data sets as was mentioned in [8]. It may be necessary to prompt the ChatGPT in a different manner to produce results that would differ from a prompt with just the addition "answer as if you were a teacher." added to the end, if the aim is to produce different associations in the answer categories as was shown in Figure 2. It is interesting to note that the ChatGPT 4.0 with the game designer prompt disagreed with the associations when compared against the other ChatGPT 4.0 data sets, which did not happen in the ChatGPT 3.5 data set.

The input collected from the ChatGPT appears to be relevant and implementable (research question 3), even though it was suggested that ChatGPT may not understand the meaning behind words [1] or take into account the context [1, 4]. As was noted by [8], no clarifying questions were asked by any of the ChatGPT versions. It may be that the prompt was lengthy enough and populated with enough relevant words to get reasonable results. Similarly to [12], as the expert participants were able to produce insight that ChatGPT did not produce, the data suggests that ChatGPT alone should not replace the need for gathering professional input for the educational game design process. The results support the idea that basing educational game design solely on theory is not as strong of a plan when compared to collaboration with educators, as was suggested by [19] or to lean solely on ChatGPT for responses as was advised in [8].

When the findings are compared to the previous research it can be said that the output of this research aligns with general views but differ in some of the suggestions for the improvement for educational games. The input from experts confirms the view mentioned in [14] that suggests that gamified elements are viewed as more attractive than non gamified, as the experts suggested adding gamification elements when they were prompted to suggest changes to the educational game if it was intended for entertainment or hobbyist purposes only. Multiplayer features and puzzles, similar to the ones discussed in [12], were found in the ChatGPT data sets. Feedback, as was mentioned in the game design principles of [10] appeared in both the expert interviews and all of the ChatGPT data sets. A clear end goal, as was discussed in [13], or a story to finish, such as was implemented in [23] are not present in any of the ChatGPT data sets. Missing from all the data sets were avatars, in game currency and trading as mentioned in [21]. It comes as no wonder that if gamification is one of the most popular topics in the reward system research corpus [16], and a trending topic in higher education [18], that both the input from expert interviews and the data sets collected from ChatGPT showed a number of suggestions that could be categorized as gamification as was described by [13]. Multitudes of the gamification elements were suggested to add motivation for the students to play the game, as mentioned in [9]'s "game motivation". In addition, the idea of adding gamification elements, according to the participants, would increase the "fun factor" of the game. This finding is opposite to the suggestion made by previous research that gamification elements are added without further thought into the purpose of it [15]. Many of the "game thinking" elements [9] such as a scoring system were present in all the data sets. These appear to support that

view in [12] that the motivation and enjoyment of the student are key values of educational game production [12].

No input that was against the touch system was collected (a finding that supports [29]). However, some input was received that favored changing the dragging of an image to the correct answer to pressing the correct answer instead.

The expert interviews seem to support the conclusion in [24] which suggests that educational games have less variety in reward types when compared to games made for enjoyment purposes only. According to the expert interviews, this may be because the educators see the learners motivation to be intrinsic and to have no need for additional motivational tools such as gamification to be used in the educational game projects.

There are multiple implications for the results of the study. The input provided by analyzing ChatGPT data sets and expert interviews be can utilized to enhance the future design and development process of educational games (research question 3). The results suggest that ChatGPT could be used as a tool to design learning tasks, as mentioned in [1], and learning materials, as mentioned in [3], in addition to producing content for game development purposes [4]. By using ChatGPT as a tool to augment the educational game development process it may be possible to reduce the workload of educational game designers. However, it is crucial to note that the participation of actual experts in addition to ChatGPT appears to be a better solution than only relying on LLM's to replace the input from experts.

4.5 Future research

The relevance of this study to the current issues in the educational game design such as the production of engaging educational games being a challenging task [10] and little empirical data on how to design educational games to meet the educational targets [11]. The use of ChatGPT is a trending topic [2], and there is a need for empirical research for the uses of ChatGPT in education [1], for which this research produces valuable input. The significance from the insight gained from this research can be used to reduce the workload of teachers as mentioned in [1].

Additional research in the use of ChatGPT as a tool to gather expert-like input is direly needed in the field. Studies that explore the use of ChatGPT and LLM's in serious and educational game design such a board games [25] and escape rooms [26] need to be expanded to include comparisons with real life experts. The usability of input from ChatGPT is an interesting field of research with many possible use cases. One interesting venue of research is to study how experts use LLM's such as ChatGPT to produce input. In addition, this field needs additional research with LLM's other than ChatGPT that compares alternative options to both expert participants and other LLM's. Furthering this field of research may lead to results that answer the questions that are asked when pondering why game elements show inconsistent results in learning as suggested by [13,14]

5. Conclusions

This study had several key questions to explore. First was to find the differences between input prompted from ChatGPT and data collected from teachers, students and game designers via a survey (research question 1). Second, to see if the addition of a persona prompt (for example, "Answer as if you were a student.") to the end of the ChatGPT prompt would produce different results (research question 2). Third, this study analyzed the data sets collected from surveys, ChatGPT version 4.0 and version 3.5 to see if the input from ChatGPT can be used in the design and development process of serious and educational games (research question 3).

The comparison of data between ChatGPT and human experts (research question 1) presented in this study has several significant implications to the development process of serious and educational games. First, the input from teachers was replicated by ChatGPT with all the prompts provided and with human game designers and students. This implies that some of the input provided for the development of serious and educational games by a group of experts may be replicated by, not only by other groups of human experts, but by any ChatGPT version when prompted with the

same query. However, there are differences between the input provided by ChatGPT 4.0 and ChatGPT 3.5.

The data sets within ChatGPT version 4.0 and ChatGPT 3.5 appear to share common associations, with the notable exception of ChatGPT 4.0 with the game designer persona prompt. The data sets between ChatGPT 3.5 and 4.0 showed different input category associations from each other. The ChatGPT 4.0 generated more input, except for ChatGPT 4.0 as a game designer, when compared to the earlier version and the categories of input are similar in frequency. In addition, ChatGPT 3.5 seems to be more resilient to changes in the prompt.

The second significant finding of this study is that the version of the ChatGPT can affect whether the input is statistically significantly similar to the input from human experts. In the case of the data set collected from human student participants it was found that all the data sets collected from ChatGPT 4.0 were statistically significantly different in comparison. Thus, when generating input that is comparable to students ChatGPT 3.5 is recommended over ChatGPT 4.0. This also shows a need for further research into ChatGPT prompts, as this finding was not replicated with other prompts used in this study.

Third, the addition of a persona prompt to the end of the ChatGPT 4.0 prompt produced statistically significantly different input when compared to input from human game designers. When ChatGPT 4.0 was prompted without a persona it produced input similar to game designer survey participants. However, the addition of a game designer persona prompt to the end of the ChatGPT 3.5 prompt did produce similar input when compared to game designer survey input. This implies that when seeking input from ChatGPT that is similar to game designers, it is recommended not to use the additional “answer as if you were a game designer” at the end of the prompt when using ChatGPT 4.0, but it is recommended when using ChatGPT 3.5. The results of this study show that the effectiveness of the addition of the persona prompt is dependent on the group of experts, ChatGPT version and the prompt used (research question 2).

The results indicate that statistically similar input can be generated for the development of educational and serious games with different ChatGPT versions and prompts and therefore ChatGPT can provide usable input for the development of serious and educational games (research question 3). The study shows that, except for ChatGPT 4.0 with the game designer persona prompt, the input categories show similar associations within the ChatGPT versions but are different from each other and the expert prompts. The data shows that it may provide useful to use both versions of ChatGPT for collecting input for the development of serious games.

In conclusion, this study shows that ChatGPT can be trusted to produce input for the development of serious games, but it shouldn’t be relied on as the only source. The results of this study have concrete impact on the use of ChatGPT for the development process of serious and educational games. The findings show that it is possible to produce statistically similar input with ChatGPT using the same prompt when compared to results from expert group surveys. However, this study also shows that the validity of this claim is heavily reliant on the prompt, the version of ChatGPT and the group of experts surveyed.

Experts produced suggestions that were unique to the expert data set when compared to all the data sets collected from the ChatGPT. A clear end point to the game and a storyline to follow were not found in any other data set. In addition, expert participants showed concerns over cheating, an input which was not proposed by ChatGPT. The study's findings suggest that ChatGPT can produce input like experts. However, it is important to note that experts were able to produce unique suggestions that did not show up in any of the ChatGPT data sets and still prove to be a necessary part of serious game design process.

The insight gained from this study can be used to leverage the use of tools such as ChatGPT 3.5 and 4.0 to reduce the workload of teachers and industry professionals and as a starting point to develop ways to harness the rapidly developing large language models in the development of serious and educational games.

Acknowledgments

The mock up used in this study was created using the Unity Game Engine. Some of the visible assets were purchased via the Unity Asset Store from the publisher "Layer Lab"

(https://assetstore.unity.com/publishers/5232). Permission for the usage of assets for the research was asked from Unity directly.

Conflicts of interest

No conflicts of interest.

References

[1] M. Farrokhnia, S. K. Banihashem, O. Noroozi & A. Wals, “A SWOT analysis of ChatGPT: Implications for educational practice and research”, *Innovations in Education and Teaching International*, 2023, pp. 1–15, doi: 10.1080/14703297.2023.2195846.
[2] R. Raman, H. Lathabhai, S. Diwakar, & P. Nedungadi, "Early research trends on ChatGPT: Insights from altmetrics and science mapping analysis," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 19, 2023, doi: 10.3991/ijet.v18i19.41793.
[3] X. Zhai, "ChatGPT and AI: The game changer for education," Available at SSRN, 2023, doi: 10.13140/RG.2.2.31107.37923.
[4] S. Biswas, "Role of ChatGPT in gaming: According to ChatGPT," Available at SSRN, 2023, doi: 10.2139/ssrn.4375510.
[5] P. L. Lanzi & D. Loiacono, "ChatGPT and other large language models as evolutionary engines for online interactive collaborative game design" arXiv preprint arXiv:2303.02155, 2023, doi: 10.1145/3583131.3590351.
[6] A. O. Ajlouni, A. S. Almahaireh, and F. A. Wahba, “Students’ perception of using ChatGPT in counseling and mental health education: The benefits and challenges,” *International Journal of Emerging Technologies in Learning*, vol. 18, no. 20, 2023, doi: 10.3991/ijet.v18i20.42075.
[7] T. T. A. Ngo, “The perception by university students of the use of ChatGPT in education,” *International Journal of Emerging Technologies in Learning*, vol. 18, no. 17, p. 4, 2023, doi: 10.3991/ijet.v18i17.39019.
[8] E. Opara, A. Mfon-Ette Theresa, and T. C. Aduke, “ChatGPT for teaching, learning and research: Prospects and challenges,” *Glob Acad J Humanit Soc Sci* vol.5, 2023, doi:10.36348/gajhss.2023.v05i02.001.
[9] J. Zeng, S. Parks, and J. Shang, “To learn scientifically, effectively, and enjoyably: A review of educational games,” Human Behavior and Emerging Technologies, vol. 2, no. 2, pp. 186–195, 2020, doi: 10.1002/hbe2.188.
[10] T. H. Laine and R. S. N. Lindberg, "Designing engaging games for education: A systematic literature review on game motivators and design principles," *IEEE Transactions on Learning Technologies*, vol. 13, no. 4, pp. 804–821, 2020, doi: 10.1109/TLT.2020.3018503.
[11] M. Oren, S. Pedersen, and K. L. Butler-Purry, "Teaching digital circuit design with a 3-D video game: The impact of using in-game tools on students’ performance," *IEEE Transactions on Education*, vol. 64, no. 1, pp. 24–31, 2020, doi: 10.1109/TE.2020.3000955.
[12] A. M. Moosa, N. Al-Maadeed, M. Saleh, S. A. Al-Maadeed, and J. M. Aljaam, "Designing a mobile serious game for raising awareness of diabetic children," *IEEE Access*, vol. 8, pp. 222876–222889, 2020, doi: 10.1109/ACCESS.2020.3043840.
[13] Z. Luo, "Gamification for educational purposes: What are the factors contributing to varied effectiveness?" *Education and Information Technologies*, vol. 27, no. 1, pp. 891–915, 2022, doi: 10.1007/s10639-021-10642-9.
[14] M. Ninaus, K. Kiili, G. Wood, K. Moeller, and S. E. Kober, "To add or not to add game elements? Exploring the effects of different cognitive task designs using eye tracking," *IEEE Transactions on Learning Technologies*, vol. 13, no. 4, pp. 847–860, 2020, doi: 10.1109/TLT.2020.3031644.
[15] D. John, N. Hussin, M. K. Zaini, D. S. Ametefe, A. A. Alju, and A. Caliskan, "Gamification Equilibrium: The Fulcrum for Balanced Intrinsic Motivation and Extrinsic Rewards in Learning Systems," *International Journal of Serious Games*, vol. 10, no. 3, pp. 83-116, Sep. 2023, doi:

10.17083/ijsg.v10i3.633.
[16] J. Tyni, A. Tarkiainen, S. López-Pernas, M. Saqr, J. Kahila, R. Bednarik, and M. Tedre, "Games and rewards: A scientometric study of rewards in educational and serious games," *IEEE Access*, vol. 10, pp. 31578–31585, 2022, doi: 10.1109/ACCESS.2022.3160230.
[17] Y. Pan and G. L. Mow, "Study on the impact of gamified teaching using mobile technology on college students' learning engagement," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 14, 2023, doi: 10.3991/ijet.v18i14.41207.
[18] I. Irwanto, D. Wahyudiati, A. D. Saputro, and S. D. Laksana, "Research trends and applications of gamification in higher education: A bibliometric analysis spanning 2013–2022," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 5, 2023, doi: 10.3991/ijet.v18i05.37021.
[19] A. Es-Sajjade and F. Paas, "Educational theories and computer game design: Lessons from an experiment in elementary mathematics education," *Educational Technology Research and Development*, vol. 68, no. 5, pp. 2685–2703, 2020, doi: 10.1007/s11423-020-09799-w.
[20] E. Pacheco-Velazquez, V. Rodes-Paragarino, L. Rabago-Mayer, and A. Bester, "How to create serious games? Proposal for a participatory methodology," *International Journal of Serious Games*, vol. 10, no. 4, pp. 55-73, Nov. 2023, doi: 10.17083/ijsg.v10i4.642.
[21] T. H. Laine, N. Duong, H. Lindvall, S. S. Oyelere, S. Rutberg, and A. K. Lindqvist, "A reusable multiplayer game for promoting active school transport: Development study," *JMIR Serious Games*, vol. 10, no. 1, e31638, 2022, doi: 10.2196/31638.
[22] S. H. Edwards and Z. Li, "A proposal to use gamification systematically to nudge students toward productive behaviors," in *Proceedings of the 20th Koli Calling International Conference on Computing Education Research*, 2020, pp. 1–8, doi: 10.1145/3428029.3428057.
[23] K. Graham, J. Anderson, C. Rife, B. Heitmeyer, P. R. Patel, S. Nykl, et al., "Cyberspace odyssey: A competitive team-oriented serious game in computer networking," *IEEE Transactions on Learning Technologies*, vol. 13, no. 3, pp. 502-515, 2020, doi: 10.1109/TLT.2020.3008607.
[24] J. Tyni, A. Turunen, J. Kahila, R. Bednarik, and M. Tedre, "Reward types in popular recreational and educational mobile games," *IEEE Access*, 2022, doi: 10.1109/ACCESS.2022.3231936.
[25] W. G. Junior, E. Marasco, B. Kim, L. Behjat, and M. Eggermont, "How ChatGPT can inspire and improve serious board game design," *International Journal of Serious Games*, vol. 10, no. 4, pp. 33-54, Nov. 2023, doi: 10.17083/ijsg.v10i4.645.
[26] P. Fotaris, T. Mastoras, and P. Lameras, "Designing educational escape rooms with generative AI: A framework and ChatGPT prompt engineering guide," in *Proceedings of the 17th European Conference on Game-Based Learning (ECGBL)*, Sep. 2023, Academic Conferences and Publishing Limited, doi: 10.34190/ecgbl.17.1.1870.
[27] H. F. Hsieh and S. E. Shannon, "Three approaches to qualitative content analysis," *Qualitative Health Research*, vol. 15, no. 9, pp. 1277–1288, 2005, doi: 10.1177/1049732305276687.
[28] S. Elo and H. Kyngäs, "The qualitative content analysis process," *Journal of Advanced Nursing*, vol. 62, no. 1, pp. 107–115, 2008, doi: 10.1111/j.1365-2648.2007.04569.x.
[29] M. Wang, "Effects of touch-type online educational games on learners' learning motivations," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 6, p. 4, 2023, doi: 10.3991/ijet.v18i06.37817.
[30] T. Wei and V. Simko, "R package 'corrplot': Visualization of a correlation matrix," Retrieved from https://github.com/taiyun/corrplot, Version 0.92, 2021.