

Generative AI in Game Pedagogy

AI Game Test Questionnaire Assistant

By

Qingyang Wang
Zhechuan Hu
Mingliang Wang

A Project Report
submitted to the Faculty
of
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Master of Science
in
Interactive Media and Game Development (IMGD)
April 25, 2024

Advisors:

Professor Gillian Smith
Professor Yunus Telli

This report represents work of WPI graduate students submitted to the faculty as evidence of a partial degree requirement. WPI routinely publishes these reports on its website without editorial or peer review. For more information about the projects program at WPI, see <http://www.wpi.edu/Academics/Projects>.

Abstract

Playtesting is essential to game development, as it is the portal between game developers and players. A reliable way to communicate feelings and feedback from playtesters can help the game development team facilitate the iterative game creation process. With the current technical development in artificial intelligence (AI), we can combine existing powerful AI functions to help students design playtesting protocols, collect data, educate, and take other considerations into account. As a proof-of-concept, we designed and developed an AI Game Questionnaire Assistant. AI Game Test Questionnaire Assistant is a customized generative AI tool built with the GPT builder function published by OpenAI. It is designed to help users create a game test questionnaire or host game testing based on their design goal. The main objective of this AI tool is to communicate with users to gather information about users' game projects through text or by analyzing uploaded documents. Then, it proposed a series of questions to better accumulate information around the context. After that, this AI agent will generate a comprehensive guide addressing the user's specific needs. To understand the intervention of the proposed AI tool, we conducted a user study to gather feedback. Our observation of the interaction between testers and AI agents examines generative AI's potential applications in game design educational settings. We iterate on the AI tool design based on the testing feedback. Our data analysis provides insights into designing and creating generative AI-based educational feedback tools.

Acknowledgment

We want to extend our heartfelt thanks to several key individuals whose support and input have been essential to the success of our project.

We are grateful for the help from Prof. Walt Yarbrough and Monty Sharma for providing their insights from the perspective of experts in the game development industry.

We are deeply grateful to Max Chen, who helped us figure out the way forward in the early project stages. She also provided us with indispensable help in the report writing stage, and it is safe to say that this report would not have been possible without her.

We are also grateful to Gillian Smith and Yunus Tellieli for their aid and counsel in our pursuit of academia. As our advisors, their expertise and mentorship have been fundamental in shaping the direction and execution of this work.

We also wish to acknowledge PhD students in our research group, Darren Cole, Josiah Boucher, and Karen Royer. Their collaboration and shared insights have greatly enriched our experience and contributed significantly to our project's development.

Our sincere thanks go to all the testers who evaluated our project. Their feedback was crucial in refining our work and enhancing its quality and effectiveness.

They have helped us, and we are profoundly thankful for their contributions.

We use AI tools to fix our grammar errors.

Table of content

1. Introduction	4
2. Background	7
2.1 Purpose of Game Testing	7
2.2 Purpose of Questionnaire	8
2.3 Difficulties in Questionnaire Design	8
2.4 Generative AI Tool	9
2.5 Problem Statement	10
2.5.1 Student Problem in Game Testing	10
2.5.2 The Concerns of Using Generative AI in Education	12
3. Design	14
3.1 Design Goal	14
3.1.1 Problem Solution	14
3.1.2 Educational Aspect	17
3.2 AI Agent Design	18
3.3 Behavior Limitation	19
4. Evaluation	21
4.1 First Prototype	21
4.1.1 General Role and Behavior	21
4.1.2 Knowledge	23
4.1.3 Solve Problems	26
4.2 First User Test	29
4.3 Second Prototype	35
4.3.1 Role Behavior Update	35
4.3.2 Introduction Update	36
4.3.3 Knowledge Update	37
4.3.4 Starting Question	37
4.4 Second User Test	43
4.5 Final Result	48
5. Conclusion	49
REFERENCE	51
APPENDICES	55
A: First-round survey	55

1. Introduction

Playtesting is essential in developing digital games, as engaging the players in the development process is essential. This process is significant because it allows game designers to identify and resolve potential issues before release. In the industry, commercial game development often involves many rounds of testing sessions. Indie developers, especially students, usually have less choice in playtesting methods due to the budget limitations and size of the developing team. Typically, independent (indie) game projects employ small-scale testing methods (Mirza-Babaei et al., 2016). One game testing method that is well-accepted is a questionnaire. Having playtesters fill out the questionnaire after playtesting sessions is an efficient and low-cost way for indie game developers and students to test their games because it can effectively measure different dimensions of player experience (Law et al., 2018). In addition, it is easy to distribute via emails or online form tools.

However, there are still problems when student game developers utilize questionnaires to hold a playtesting session. To further understand the challenges around designing and conducting playtesting sessions, we interviewed game industry veteran Prof. Walt Yarbrough. He mentioned two main challenges for students: unawareness of the objective of playtesting and the “one size fits all” playtesting design. Many students, when designing the playtesting session for their game creation, are often confused by the purpose of doing playtesting. That results in the fact that they tend to ask questions to reassure them that their games are good. The other challenge is that students’ playtesting procedure often aims to fit the majority, which means their playtesting designs are similar. So, students cannot take advantage of playtesting to reflect various dimensions of their games. However, providing individual feedback based on the condition concerning the specific genre of the game requires a tremendous amount of effort from the instructor or mentor of the team.

Therefore, we proposed our research question, "How can we guide students with individual feedback so that they can iteratively design their game testing processes independently?". To address this question, we developed an AI tool to assist students in creating playtesting questionnaires. It is a gateway for students better to understand game testing and their personal game projects. After reading academic articles regarding game testing, we became interested in the automation concept. Politowski et al. (2021) suggest some game companies have already employed deep-learning algorithms to develop automated game testing agents. In this case, automation is still designed for certain game testing, but generative AI expands the possibility of providing individual feedback for all students in designing game testing.

AI techniques have developed rapidly in recent years and have demonstrated powerful functions to their users with Large Language Models (LLMs). Based on the Generative Pre-Trained Transformer (GPT) technique published by OpenAI, the AI chatbot ChatGPT can generate real-time feedback based on the user's prompt (OpenAI, 2024). We aim to design a real-time interaction method to simulate the teaching environment between teachers and students. Student game developers can have their virtual tutor to guide them in designing game test questionnaires. A recent study indicates that ChatGPT can provide personalized feedback based on the student's needs and learning progress (Baidoo-Anu & Owusu Ansah, 2023). Therefore, our project is focused on developing a GPT-based AI assistant for students to design questionnaires for playtesting.

The design section demonstrates how we create a framework derived from our research on user experience, game testing, and generative AI. The project methodology section explains the design goals and how we customized the AI assistant in ChatGPT. The evaluation section provides the results of user testing, including graphs and objective

feedback from survey responses. The conclusion includes our analysis of tester feedback and how we can further improve our AI tools.

2. Background

In this part, the project team describes the background of the project, including game testing methods, the purpose and advantages of game questionnaires, difficulties in questionnaire design, and problems encountered by generative AI in educational environments.

2.1 Purpose of Game Testing

Game testing is an important game development activity because it allows developers to uncover interaction problems (Redavid et al., 2011). The development and refinement of video games is a nuanced process that can get significant improvement from the feedback of playtesting. As a form of software development, this process ensures that the objectives of the game project are met within the allocated budget and released on schedule, fulfilling the specified requirements and achieving acceptable quality (Kasurinen & Smolander, 2014). Playtests serve a significant function by helping game designers verify whether the game is time manageable and meets the original goals. This essential process identifies a game's strengths and weaknesses and allows developers to make necessary adjustments and enhancements.

A comprehensive game development project typically includes multiple playtests across different development stages. Depending on the specific goal, each playtest helps the developer revise the project and guides it in diverse potential developmental directions.

Game testing is also a requirement in students' projects. However, the lack of a clear framework and related knowledge of designing a game test challenges students in figuring out their test goals. Thus, students may encounter many problems during the game test and feedback analysis.

2.2 Purpose of Questionnaire

The questionnaire is an important tool for collecting information from people in various application scenarios (Steinmaurer et al., 2021). It is a common way for students to collect feedback from the playtest, determining testers' gaming experiences and the direction for improvements. Students can obtain consistent and quantifiable data through the reasonable organization of questions in their questionnaires. Even though creating a good questionnaire can be time-consuming, designing and deploying a high-quality questionnaire in the playtest is extremely valuable.

Due to its straightforward format and quick editability, the questionnaire is widely used in individual and student game projects, making it more important to study and research. In this context, a related AI tool that advises students could significantly reduce their time on collecting versatile feedback from testers.

2.3 Difficulties in Questionnaire Design

A good question must correspond to a particular need and reflect clear goals (Bhandari, 2023), and a solid questionnaire is made up of good questions, which means there is no general template to use directly for the questionnaire. So, designing a questionnaire requires extensive preparatory work. The design process involves creating a detailed conceptual framework, selecting appropriate question types (open-ended or closed-ended), and ensuring clarity and relevance in the questions to meet the objectives. As a result, students must spend more time practicing literature skills, like defining concepts and constructive writing, in designing a questionnaire. Additionally, questions should be tested for reliability and validity, which often requires testing with feedback adjustments to guarantee that the questionnaire accurately captures the intended data without bias.

Differences in conceptual framework, choices of question types, and clarity of the bond between questions and objectives can result in considerable differences in the quality of the final product. Without explicit instruction or practicing, students may spend plenty of time on questionnaire design but get a low-quality product finally. Besides this, unclear questionnaires may complicate the data analysis process and present more difficulties for students.

2.4 Generative AI Tool

Since OpenAI company published its Generative Artificial Intelligence (GAI) tool, ChatGPT 4.0, which is based on the Large Language Model (LLM), more and more people have realized the capability and user-friendliness of such tools. During the same period, various other GAI tools targeting different audiences have emerged on the market, as well as expanding the game industry's possibilities. These GAI tools have a disruptive influence on game testing: with LLM, these advanced AI techniques can comprehend users' questions and text descriptions and generate precise feedback or summary for the input prompt (Aydın & Karaarslan, 2023). With these capabilities, generative AI tools can gather information related to users' game projects from text descriptions, then it can design corresponding questionnaires for users.

Although Generative AI technologies have been used extensively in commercial companies or cutting-edge research institutions, their potential to analyze small game projects and give advice on questionnaire design is still underestimated. Therefore, it is worth using AI tools to optimize the questionnaire creation process and enhance efficiency for individuals like student game designers.

2.5 Problem Statement

Before creating the AI tool, we would like to learn about different challenges that students are facing in terms of designing playtesting, especially in designing questionnaires. In this part, we will discuss several problems found when consulting the game industry veteran.

2.5.1 Student Problem in Game Testing

After establishing that our AI tool would primarily focus on playtesting for students and independent game makers, we interviewed Prof. Walt Yarbrough.

Prof. Yarbrough is an experienced game management practitioner who has worked in the game industry for over ten years. His experience as a project manager and professor provides him with unique perspectives on game testing in the industry and student-created games. Prof. Yarbrough noticed three challenges in playtesting when he was advising a group of students in designing a website for scheduling game playtesting. He mentioned students often started their playtest too late. Playtest is essential to help designers detect problems early (Pietriková & Sobota, 2023). Students always host playtesting sessions after finishing their game, but the right time for a playtest will be during development time. Then, we discussed the three most significant problems in the student playtesting session.

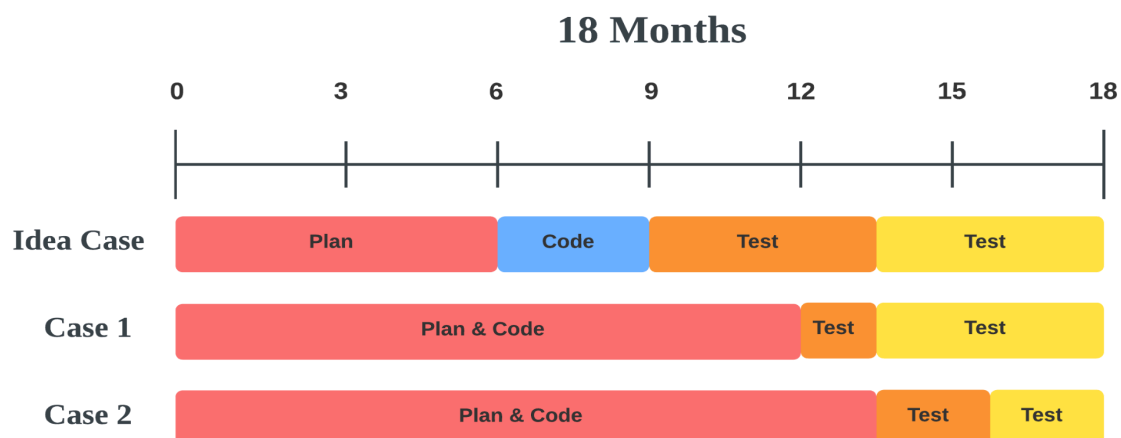


Figure 1. The time game test should be hosted.

- **Overly Positive Feedback**

A prevalent issue within educational settings, particularly among students venturing into game development, is the reception of overly positive feedback. Because the creator is the one who makes the game, it is hard to feel comfortable hearing someone criticize it (Lewis-Evans, 2023). People are affected by social stress. They would not tell their true thoughts to the game designer; they want to please the designer, which will make the game unable to improve. Therefore, testers frequently award high ratings, such as 8 or 9 out of 10, which will inaccurately reflect the game's areas in need of improvement. This trend of non-critical evaluation is counterproductive, as it stifles the potential for iterative refinement—a cornerstone in game development.

- **Inadequate Question Design Leading to Superficial Feedback**

The questions posed to testers often do not facilitate deep, constructive criticism but rather encourage superficial praise. Good questions could lead to more insightful feedback, such as inquiries about the game's completion status, engagement level, and replay value, but they are seldom asked by students in their questionnaires. Without comprehensive feedback, the game is just not as good as it could be through the round-by-round game testing. So, teaching students to design questionnaires will be vital for them in improving their game. A helpful lesson can include questions like “what type of game testing they want to use”, “to what scale they want to use”, and “what questions they should ask” for students. The student project team should draw these questions to help them to design a helpful questionnaire.

- **Misunderstanding the Purpose of Playtesting**

Effective user testing could test art or usability, as well as game rules and mechanics (Isbister & Schaffer, 2008). Many students fail to recognize playtesting as a critical improvement opportunity, seeing it as a mere formality. This misperception extends to misconceptions about market needs and the commercial viability of their games. For those

aiming to sell their game, understanding customer willingness to pay is essential yet often overlooked. Similarly, students who are looking to enter the game industry might not fully recognize a few professional feedback from peer evaluations. Therefore, for some students who want to sell their games, how to ask testers if they would be willing to buy their game will be crucial. For some students who just made a game for their portfolio, getting enough valuable suggestions from professionals will be vital for them to improve their portfolio and find a good job. However, the real situation we found is that students don't know the goals of testing in their games, so they often cannot host a game testing well. In addition, students feel that game testing is a task that needs to be completed to earn credit in class instead of an effective way to improve their games.

2.5.2 The Concerns of Using Generative AI in Education

The sudden rise of generative AI, especially the ChatGPT published by OpenAI, has greatly impacted the field of education. Some people think it could enhance students' study by aiding in preparing for the class (Albert & Li, 2023). Although generative AI has many advantages in education, there are still problems here. For example, many concerns in the academic field come from the fact that using ChatGPT to generate papers, such as non-existent citations and irrelevant references, will bring great confusion (Stokel-Walker, 2022). Students use the content generated by chatbots to complete assignments and exams, and it is hard for instructors to distinguish the generated content from human-writing one. Although ChatGPT can achieve high scores in certain aspects of exams, it harms the academic integrity in class. In many educators' views, this type of artificial intelligence has originality, plagiarism, and ethical issues. (Chatterjee & Dethlefs, 2023; Stokel-Walker, 2022). However, ChatGPT's capabilities are obvious to all. When GPT comes to searching and summarizing resources, it is considered to be in good agreement with those who are competent and generally exhibit the same characteristics (e.g., efficient, skilled, and

resourceful) (Ferraro et al., 2023; Rychen & Salganik, 2003). Therefore, ChatGPT users can obtain efficient advice from their interactions with it. This opinion motivated us to connect the Generative AI tools with game design pedagogy.

3. Design

Our customized AI agent, called “AI Game Design Questionnaire Assistant”, is built upon the GPT builder function. All foundational design logic follows the official instruction document published by OpenAI. With the official instructions, the project team combined the “Role and Goal,” “Constraints,” “Clarification,” and “Personalization” together (*GPT Builder* | *OpenAI Help Center. n.d.*). These four aspects cover behavior patterns, response tones, behavior limitations, and the specific requirements of the customized AI agent.

Besides the essential role framework and instructions, the project team also explored the possibility of controlling the AI agent’s behavior in the design process.

3.1 Design Goal

Following the interview, we first summarized the students' problems: getting overrated results from game testing, using superficial questions in questionnaires, and having unclear goals in game testing. Targeting these problems, the project team planned to design a questionnaire assistant to generate high-quality questionnaires for users. Besides this, the project team also tried to implement the education goal of this AI tool, which can inspire users’ study process during the interacting conversation with the AI agent. This goal also led the project team to design specific instructions to set limitations of the AI agent’s behavior. The project team wished to provide a more engaging experience for users with this method.

3.1.1 Problem Solution

- **Allowing testers to give honest answers can solve the problem of overly positive reviews:**

Why people give students such high ratings should be figured out to address the issue of over-positive feedback. There are three reasons for people exaggerating on the playtest

questionnaire: social-desirability bias, self-censorship, and preference falsification.

Social-desirability bias happens when people respond to questions in a way they think is more acceptable or favorable than honestly (Krumpal, 2011b). Self-censorship is the act of censoring or classifying one's discourse, usually due to fear of social pressure or to conform to social rules (Wikipedia contributors, 2024b). Preference Falsification means misrepresenting one's genuine wants, needs, or opinions to conform to what is perceived as socially acceptable or beneficial (Wikipedia contributors, 2024b). The situation is that people do not want to give the game an honest review due to all sorts of mental pressures, even if the game does not deserve a high grade. Responses to sensitive questions are often influenced by social desirability bias (Krumpal, 2013). To mitigate these mental pressures, it is important to eschew bias-inducing statements. Creating neutral questions could help a lot in the context of game-testing questionnaires. For example, rather than positing, "Most people find this game exciting," a more neutral phrasing such as "How exciting do you find this game?" should be better. This methodology promotes honest feedback, which is essential for acquiring accurate data in game testing.

Also, it should be recommended that Likert scale is used to develop questionnaires. Compared with the YES or NO binary scale, the Likert scale performs better regarding sensitivity because it has more dimensions, such as "Strongly Agree" to "Strongly Disagree." This multi-choice scale allows test takers to express their attitudes more freely in a neutral attitude that cannot be reflected on the YES/NO scale. Therefore, our design adopts a mixed questionnaire structure.

Truthful feedback directly contributes to the accuracy of research findings and can facilitate substantive improvements in gameplay, which is essential. Other methods can also increase the probability of getting accurate answers from test takers, such as anonymity,

asking respondents to answer truthfully, and test administration (Korb, 2011). These methods inform participants about the importance of honest input and encourage them to participate in the questionnaire in good faith to increase the validity of the data. Since it is impossible to put all the papers on dishonest action into the artificial intelligence system database, the project team first collected these papers on how to solve the problems of social desirability bias, self-censorship, and preference falsification. Then, it summarized the solutions to create a Txt file to help chatbots use this content to alert users to these issues and tell them solutions.

- **Use assistance to help users create questions:**

As for some students and indie game makers who cannot ask good questions, the project team has also made great efforts to help them. We use GUESS (Game User Experience Satisfaction Scale) as a framework to guide users in designing questionnaires. The GUESS is a psychometrically validated and comprehensive gaming scale with nine subscales (e.g., Usability/Playability, Creative Freedom, and Social Connectivity) (Phan et al., 2016). It has been tested on hundreds of games and is pretty reliable. As a famous framework, it can guide students in designing the questionnaire. It can provide many questions about how to judge whether players are satisfied with the game content.

In addition, because AI systems have many different questionnaires from different games as a reference, they could output more valuable content.

- **Teach users the importance of testing and how to achieve their goals:**

The misunderstanding of game testing among students is a more severe problem. Many students regard game testing only as homework that should be completed and a process that can be skipped. The point is that they have no idea of the purpose of playtesting.

The project team summarized the three purposes of students making games through interviews with professors: business, portfolio, and skill learning. For that different purpose, specific questions will be precious. For the student who wants to know their game's market

value, the suggestion will be to ask their tester if they want to spend money to buy their game, and what price they could accept will be very important. Moreover, by suggesting that they compare their game with some famous game, it will be much easier to make them find their game's position and ensure how much the customer will pay. For someone who made a game just for a portfolio and training skills, seeking feedback from industry professionals such as game design professors or experienced developers will be good. This input is invaluable for improving game design and technical skills. The project team will try an AI system to help students avoid these problems.

3.1.2 Educational Aspect

The GPT model uses much digital content data and is based on natural language processing (NLP). It can comprehend the user's textual prompts and generate logical and usable responses. Some studies have shown that AI-generated text is similar to human-made text (Aydin & Karaarslan, 2023). These powerful abilities offer ChatGPT the possibility to work in the education zone.

To be a good assistant, the AI agent must comprehensively analyze the user's game project and game test. Based on this function, the AI agent should consider the education goal to realize the learning process. The first design concept is to ask students more questions. The AI agent can gather more details from the user's description when asking questions. In answering questions, the users can recall their projects for further thought and reflection. This process shows that more questions from the AI agent can provide two-way positive feedback.

Another concept is to avoid providing direct answers to users. ChatGPT was designed to create the questionnaire directly, significantly weakening education's significance. Students may give up thinking about their questions due to the straight answers. Therefore, the project team customized the AI agent to generate a guide rather than a complete questionnaire. This

approach strengthens the learning process during the conversation and provides users with clear guidance for quickly formulating the final questionnaire.

3.2 AI Agent Design

The project team commits to creating a more “human-like” conversation experience. The first part is the “Clarification,” which ensures the responses and generations from this AI agent are comprehensible and engaging. The whole conversation will focus on the initial topic and move to the next one after thoroughly discussing it. The AI agent's answer should be brief and valuable rather than listing bullet points.

The GPT builder function offers developers two options for more detailed settings. The first method is interacting with the AI assistant (Figure 2), which can modify the AI agent directly. Developers can specify their requirements, and the assistant will adjust the AI agent automatically based on the input text prompts.

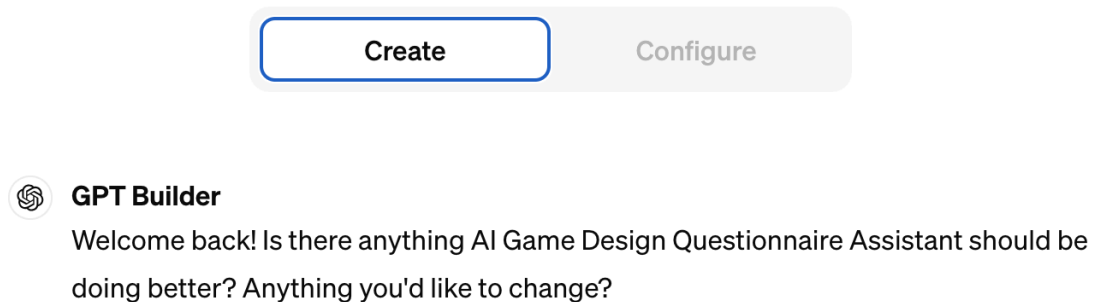


Figure 2. GPT builder interface

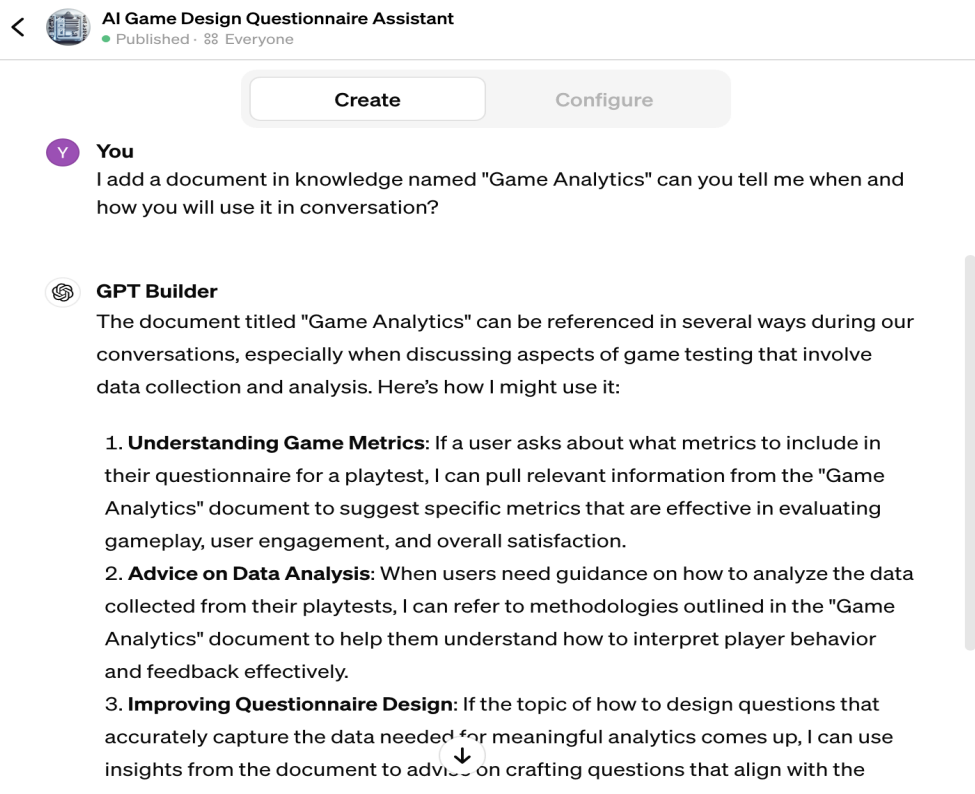


Figure 3. Paper Testing

Another method is to write the specific instructions to tell the GPT builder how to use training data sets which we provide in the configuration section. Compared with the first method, this is more time-consuming but allows developers to implement more details. The project team experimented with both methods in the design process and evaluated the final results. To provide detailed and precise feedback, the project team used handwriting to build the whole behavior pattern. In addition, we also found that GPT builder is suitable for examining training data sets and simulating the usage environment for specific documents. So, we can often check in with GPT to make sure the use of training data sets is on the right track.

3.3 Behavior Limitation

Corresponding to the “human-like” conversation concept, strict behavior control is essential and inevitable. The specific direction will be based on the two primary design goals.

To facilitate the learning process, the AI agent is required to raise questions to gather more information. So, the first limitation is to force the AI agent to ask questions. The project team also designed several behavior pattern instructions to test the effect. The verbose feedback distracts the user and makes the conversation boring. Therefore, another strategy is to limit the length of responses and ensure the conversation is concise and informative. Users can get direct answers and straightforward questions during the interaction to help them engage in the study process.

The design process determined the direction and general research methods, and with constant experimentation and comparison, the project team finally got a straightforward approach for the AI agent. The project team designed several step-by-step instructions to create behavior patterns for the AI agent.

4. Evaluation

The project team used several methods and approaches to build customizing generative content, behavior-controlling, and user guidance to achieve the expected effect in the design process. Then, the project team hosted several user tests to examine how to enhance the learning process. After each user test, the project team analyzed data and modified the existing approach to achieve better effects.

“GPTs are custom versions of ChatGPT that users can tailor for specific tasks or topics by combining instructions, knowledge, and capabilities. They can be as simple or complex as needed, addressing anything from language learning to technical support”
(GPTS FAQ | OpenAI Help Center, n.d.).

We use this tool published by OpenAI to customize the AI agent.

4.1 First Prototype

After the design process, we completed several settings in the AI agent, including basic role descriptions, general descriptions, and training data sets.

4.1.1 General Role and Behavior

This part is about the specific method used in the project to shape the AI agent’s general role and behaviors. This process follows the description from the official GPT builder guide, which covers the role description and introduces general behavior.

Role description

In the customized AI agent, role descriptions can give the GPT builder a unique speaking tone. The different tones may affect the conversation experience significantly.

The project team gave the AI agent an assistant status and asked the AI agent to maintain a professional tone throughout the conversation. We wished to create an educational

atmosphere in conversation with a positive tone. Below is the initial description used in the project:

“As the AI Game Design Questionnaire Assistant, your primary focus should be understanding and addressing user queries regarding game design questionnaires in a step-by-step manner.”

In the initial settings, all role descriptions used the second-person perspective. The project team added the step-by-step behavior limitation within the role description related to the behavior control exploration. In this setting, our priority focused on the “assistant,” “questionnaires,” and “step-by-step.” The project team intends to use this approach to break the extended response into several small pieces.

General Introduction

The description in the “instruction” part determines the behavior pattern of the AI agent. The project team will write behavior instructions in this section.

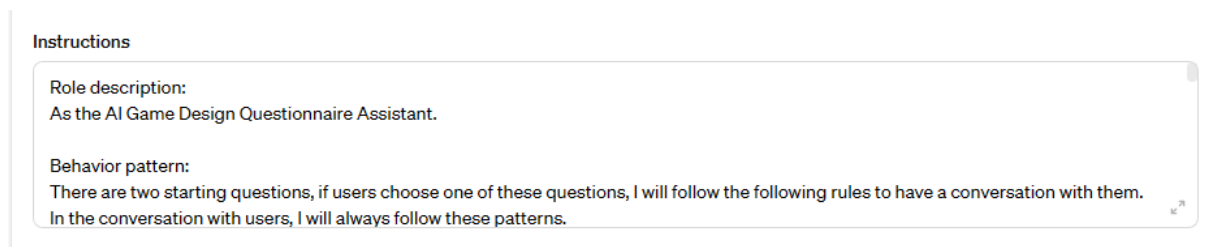


Figure 4. Instruction interface

As one of the design goals, the project team intends to set up substantial behavior limitations on the AI agent’s response and answer pattern. The initial attempt was to set behavior patterns in the description role (as the example in the last section). Corresponding to the step-by-step instruction there, we wrote specific behaviors in each step (Form 1).

As the AI Game Design Questionnaire Assistant, your primary focus should be understanding and addressing user queries regarding game design questionnaires in a step-by-step manner. Here are several steps you must follow:

1. Ask questions to the user about their "survey goal". Especially the goal of their game

project, if they want to put that into the portfolio, for commercial use, or for educational use.

2. This step is very important. After users answer these questions, carefully dissect them to understand each aspect thoroughly and ask for more details until you get enough information. After getting enough information, reclaim and summarize it first. And based on the training dataset, you need to check if the user's answer is good to use. If their answer is not good, you need to require them to answer again. Every answer and question must follow this step and standard.
3. Ask "Target Audience" to users and also follow the pattern and standard of step 2.
4. Ask "The length and details of the game test" to users and also follow the pattern and standard of step 2.
5. Ask "specific data" to users and also follow the pattern and standard of step 2.
6. Ask "Existing Outline or Content" to users and also follow the pattern and standard of step 2.
7. Give help in providing comprehensive and accurate assistance for users. Keep your responses focused on helping users create effective questionnaires for game design, offering insights based on principles of game testing and questionnaire design. Remember to maintain an informative and helpful tone throughout the interaction to facilitate the user's understanding of game design and questionnaire development.
8. You do not need to generate a questionnaire for them but must list an outline for their game test and questionnaire making. And if that is possible, ask questions to users to make sure they have understood and learned from this chat.

Form 1: First version of Instruction

In this instruction, we wrote eight steps to shape an AI agent's behavior. In most steps, we used "ask" directly to raise more questions. The first six steps form the data collection phase. After this phase, the AI agent will give corresponding feedback on users' game projects. In step eight, we requested the AI agent not generate a completed questionnaire, but a user guideline. We tried to enhance the learning process with this method.

4.1.2 Knowledge

Developers can upload files containing additional context to GPT's knowledge functionality. After uploading these files, GPT can then access this data using various methods based on user prompts (*Knowledge in GPTs | OpenAI Help Center*, n.d.). Although the chatbot could output much content, if users want the chatbot to output unique content, the Chatbot needs to search for its "knowledge". For example, knowledge is like the difference between common sense and domain-specific knowledge.

Training Set Category

The importance and relevance of documents used in the research are categorized into three classes.

Knowledge

If you upload files under Knowledge, conversations with your GPT may include file contents. Files can be downloaded when Code Interpreter is enabled

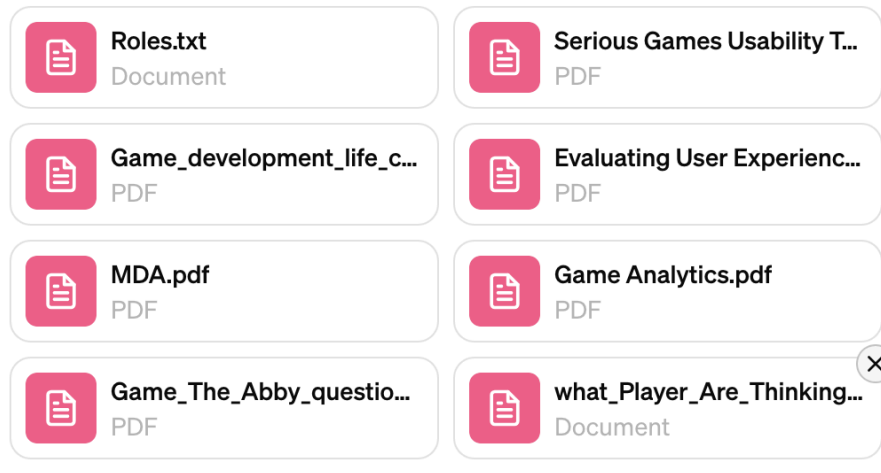


Figure 5: Training data set

- **A-class** documents are the most important and include foundational resources such as paradigms of questionnaire design, frameworks for creating questions, and scholarly papers on game testing.
- **B-class** documents, which are of secondary importance, consist of design patterns of various games and studies concerning the reliability of questionnaires.
- **C-class** documents serve as supplementary materials and typically include self-generated text files, interviews with professors, comments from internet forums, and examples of questionnaires from various games, all defined as C-class documents.

Details

After completing the design phase, the project team added some training sets in the AI agent; for the A-class training set, we added the GUESS (Phan et al., 2016) framework and some questionnaire design papers about how to design a good questionnaire. In the initial prototype, the project team did not add any B-class training sets. The problems in students we found were added to the C-class training set, and the solution document was added.

In all types of training sets, C-class documents are the most important. It stands for the private data in our customized, which will play a significant role in differentiating our AI agent from the general ChatGPT.

The first step of creating a text file is to collect necessary data, including comments from forums, information from interviews, and summaries from academic papers. Then, split them into different files based on content, like question examples or questionnaires for each game from the online player community. This data is organized into distinct text files, each labeled according to its role within the research. This categorization and labeling facilitate systematic analysis and ensure that each dataset can be efficiently accessed and utilized for specific research purposes. In general, marks tell GPT what the data is about and how it should be used.

The project team hopes to help users recognize the problem's existence with the specific output and enhance their learning process by identifying and solving problems.

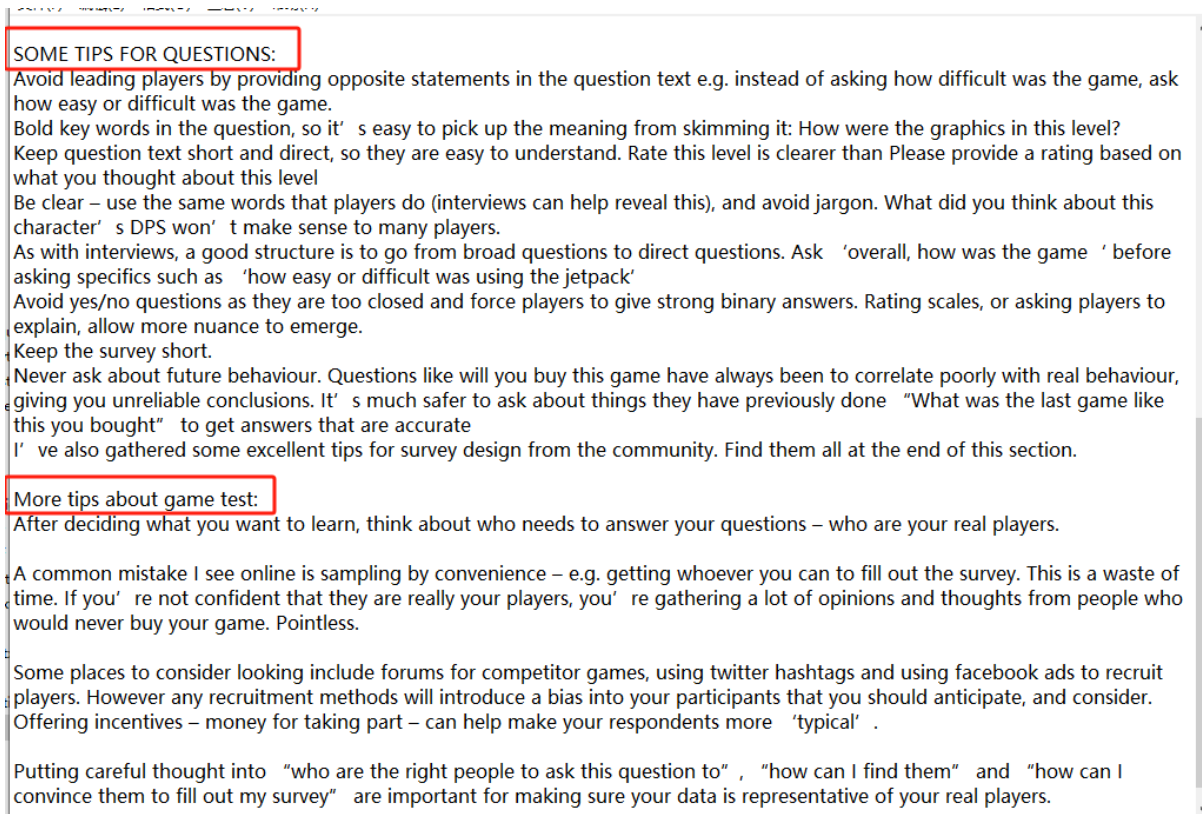


Figure 6. How to mark in-text

4.1.3 Solve Problems

In the previous content, the project team mentioned several problems students encountered and created special documentation to help students identify and solve problems. In this section, we would like to list three problems encountered by students, namely: overly positive feedback, inadequate questions leading to superficial feedback, and misunderstanding the purpose of playtesting. So, this section will focus on creating special documentation to help them identify and solve problems.

To solve the problem of over-evaluation, the project team specially wrote a text file for it (Figure 7). The content of the picture includes why people cannot answer questions honestly and the accompanying solutions. After adding this C-class file, the chatbot can remind users to use neutral language and other suggestions when creating questionnaires to

help users recognize the existence of this problem and then help users solve it.

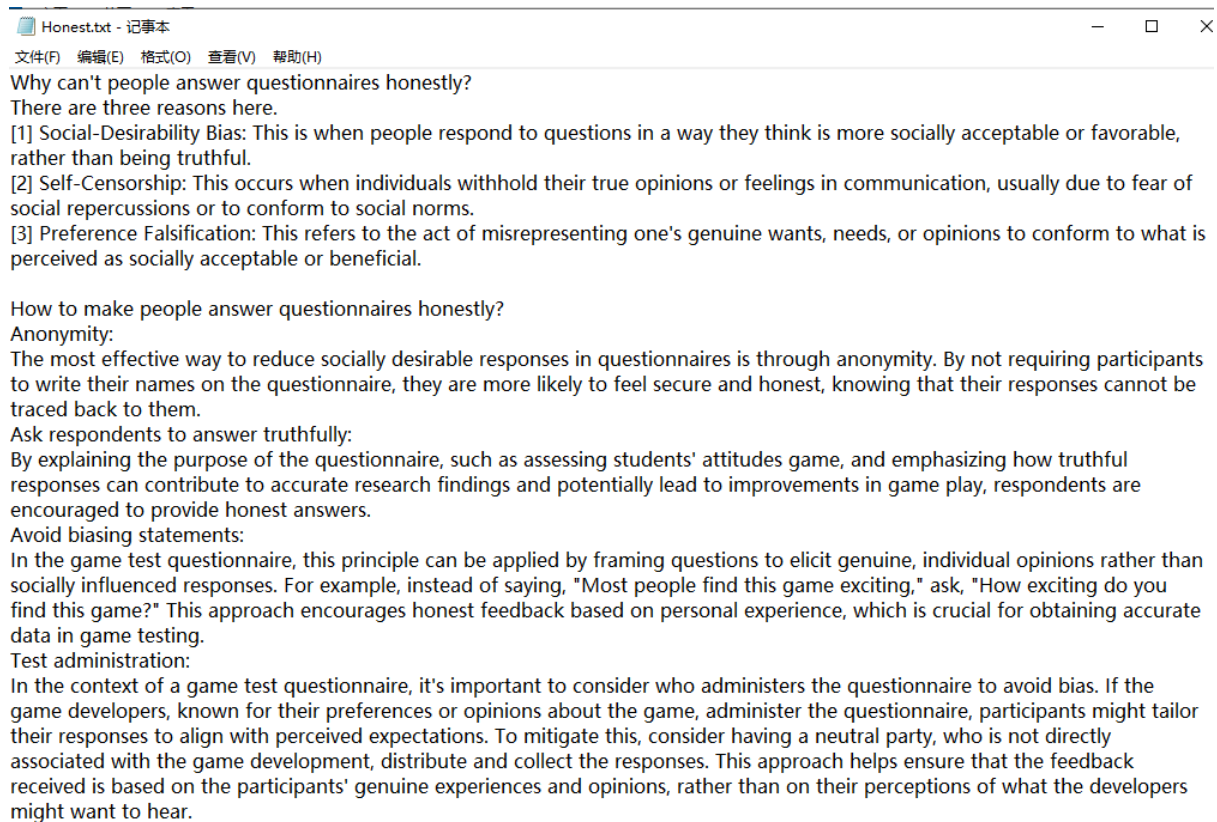


Figure 7 Special text file, to solve the dishonest problem

As for the second question, as mentioned before, the project team used GUESS as a framework to guide players in creating high-quality questionnaires to compare invalid questions. In addition to GUESS, the project team also uses game design articles and C-level files to help players create high-quality questionnaires. These files can focus on outputting the playing experience of different games to help players understand their games more deeply and ask deeper questions. As Figures 8 and 9 show, this C-class contains some sample questions and techniques for asking questions, which can help players ask more in-depth and

effective questions.

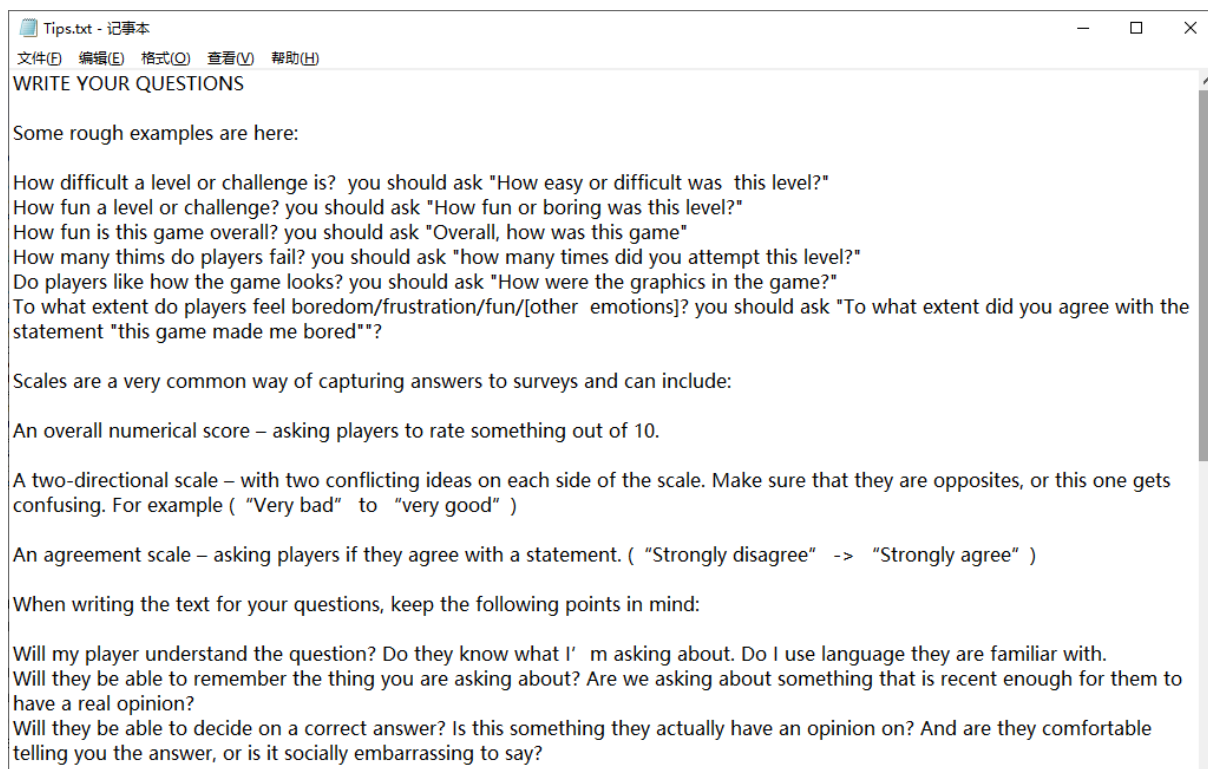


Figure 8 Special text file, to solve the second problem

SOME TIPS FOR QUESTIONS:

Avoid leading players by providing opposite statements in the question text e.g. instead of asking how difficult was the game, ask how easy or difficult was the game.

Bold key words in the question, so it' s easy to pick up the meaning from skimming it: How were the graphics in this level?

Keep question text short and direct, so they are easy to understand. Rate this level is clearer than Please provide a rating based on what you thought about this level

Be clear – use the same words that players do (interviews can help reveal this), and avoid jargon. What did you think about this character' s DPS won' t make sense to many players.

As with interviews, a good structure is to go from broad questions to direct questions. Ask 'overall, how was the game ' before asking specifics such as 'how easy or difficult was using the jetpack'

Avoid yes/no questions as they are too closed and force players to give strong binary answers. Rating scales, or asking players to explain, allow more nuance to emerge.

Keep the survey short.

Never ask about future behaviour. Questions like will you buy this game have always been to correlate poorly with real behaviour, giving you unreliable conclusions. It' s much safer to ask about things they have previously done "What was the last game like this you bought" to get answers that are accurate

I' ve also gathered some excellent tips for survey design from the community. Find them all at the end of this section.

More tips about game test:
After deciding what you want to learn, think about who needs to answer your questions – who are your real players.

A common mistake I see online is sampling by convenience – e.g. getting whoever you can to fill out the survey. This is a waste of time. If you' re not confident that they are really your players, you' re gathering a lot of opinions and thoughts from people who would never buy your game. Pointless.

Some places to consider looking include forums for competitor games, using twitter hashtags and using facebook ads to recruit players. However any recruitment methods will introduce a bias into your participants that you should anticipate, and consider. Offering incentives – money for taking part – can help make your respondents more "typical"

Figure 9 Special text file, to solve the second problem

As for the third question, a different approach was used. The project team edited the chatbot's behavior pattern to ask questions about the players' testing purposes to make players think about their testing purposes. Then it used particular text files to output suggestions to

help players complete their testing goals. Also to enable the chatbot to output unique and valuable insights on question three, the project team prepared a C-class file for it (Figure 10).

3. Misunderstanding the Purpose of Playtesting: Many students do not fully grasp the importance or objectives of playtesting. Playtesting is crucial for gathering feedback, identifying bugs, and understanding player experience, but if students see it merely as a formality or a step to complete, they may not use it effectively to improve their games. For students aiming to sell their games, there's a gap in understanding market needs and customer willingness to pay. They might be more focused on receiving positive feedback rather than testing the game's commercial viability. For students focusing on getting hired, the approach of only showcasing work within their peer group is limited.

Solution: Asking potential customers if they would pay for the game is a crucial step in validating its market potential, which these students might be overlooking. Following a model like Rockstar's, where students show their work directly to potential employers, can provide more realistic and constructive feedback. This approach exposes students to industry standards and expectations, and it offers a better assessment of how their work might be valued in a professional setting.

Figure 10 Special text file, to solve the third problem

4.2 First User Test

After completing the initial demo, the project team started the first recruitment for user tests. The target group of testers includes all Worcester Polytechnic Institution students in Interactive Media and Game Design (IMGD) majors and other related majors. Testers' academic degrees range from undergraduate to PhD. The testing form is adaptable and includes online and offline testing. For the offline session, a project team member will supply a specific GPT-4 premium account to the tester and document their interaction process. Online testers will be assigned a specific time slot to use the provided account, and their dialogues with the AI agent will also be recorded. In the test, testers need to interact with the AI agent and then fill out a questionnaire about their experience, background, and potential advice.

The first test lasted one week, and there were four testers in this test, all of whom had previous experience with generative AI tools. In the first user test, we wanted to confirm the usability of the tool, we collected data and analyzed these aspects: time cost, comprehension, final generations, and education potential.

The project team has made many efforts to avoid the dishonesty problem mentioned by Professor Yarbrough before. First, the team used a neutral tone to express the question and gave testers a Likert scale so that they could describe their thoughts as accurately as possible.

In this conversation, this tool gives you new ideas or understanding about your game projects.

1 2 3 4 5

Strongly Disagree Strongly Agree

Figure11 Examples of Likert questionnaires and questions used by the project team

To encourage testers to express their true thoughts, the project team will explain before testers participate in the test that real data can help the team improve the quality of the product before testing. And tell them that the project team will also protect the data security of all testers. In addition, to avoid the psychological pressure on the testers to please the experimenters, which would harm the experimental results, the project team used email to send the questionnaires and avoid this pressure by not having direct contact with the testers.

Benefits to Research Participants and Others: The researcher will obtain survey and interview results that will contribute to the quality and depth of this project.

Record Keeping and Confidentiality: Records of your participation in this study will be held confidential so far as permitted by law. However, the study investigators, the sponsor or its designee and, under certain circumstances, the Worcester Polytechnic Institute Institutional Review Board (WPI IRB) will be able to inspect and have access to confidential data that identify you by name. Any publication or presentation of the data will not identify you.”

Figure 12 The project team explains that honest feedback can help improve product quality and inform testers that their data is protected

Data Analysis

We designed several questions to test approaches in the initial prototype. In addition to test methods, we focused on the usability and product quality of the AI agent. These

aspects included time efficiency, comprehension, final generation quality, and educational potential.

- **Time Cost**

In this user test, the first objective is to determine the time efficiency of the AI agent. So, the first two questions examine this point by asking users about the time taken to get the final product and their experience on time-saving. According to the data feedback (Figure 13), all users save time with this AI agent, and one tester gets remarkable results. This result meets the team’s expectation of generative AI tools on time efficiency. The customized AI agent's result verifies its ability to generate the final product rapidly.

Do you think this tool saves you time compared to making a questionnaire by yourself?
4 responses

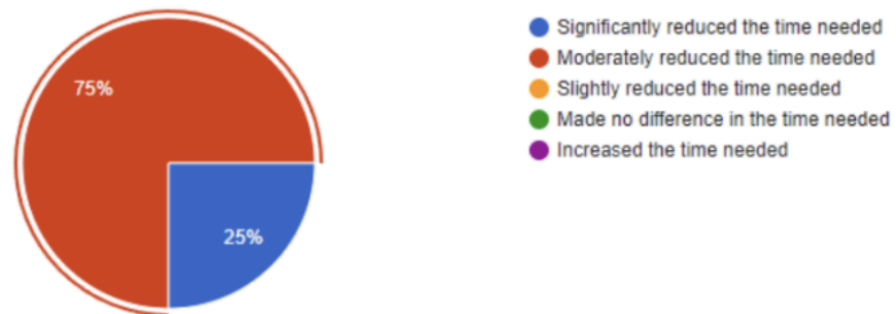


Figure 13: Time efficiency report

- **Comprehension**

This part is designed to determine if this AI agent can understand the user’s game project through only textual description or provided game-related documents (Figure 14). In this question, all users give four scores out of five, which displays that the AI tool can construct a general model of the user’s project in the conversation, but this model is still missing details.

During your conversation with this AI agent, can it clearly understand your questions and needs?

4 responses

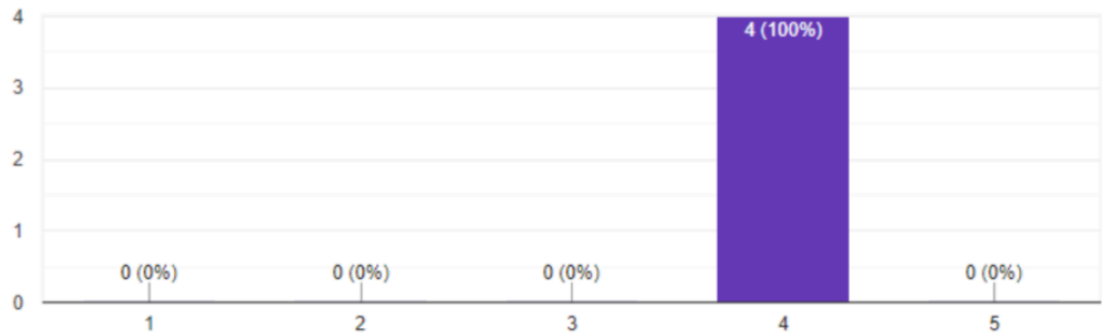


Figure 14: Comprehension reports

Another question aims to determine the user's satisfaction with the feedback. There are five options to ask how many accurate answers they get from the conversation (Figure 15). We got positive feedback from all testers, which can be seen as a possible indication of high user satisfaction.

Out of the questions you asked the AI agent, how many received accurate and satisfactory answers?

4 responses

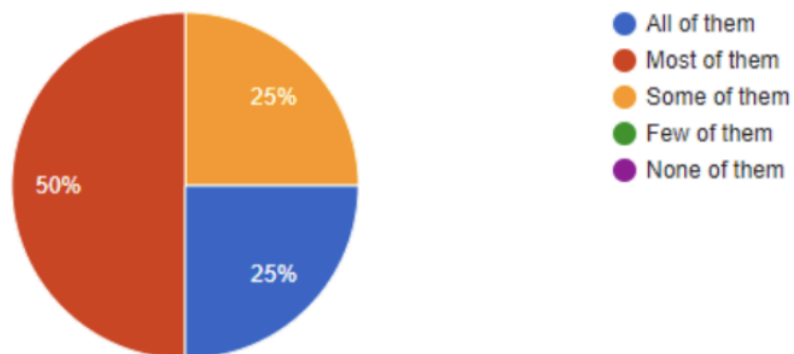


Figure 15: Accuracy report

- **Final Generations**

An essential part of the user test is to confirm the availability and quality of the final generation. The first question directly asks user satisfaction (Figure 16). In this question, all feedback falls in the range of three to five. The quality of the final product is acceptable, but there is still potential for enhancement.

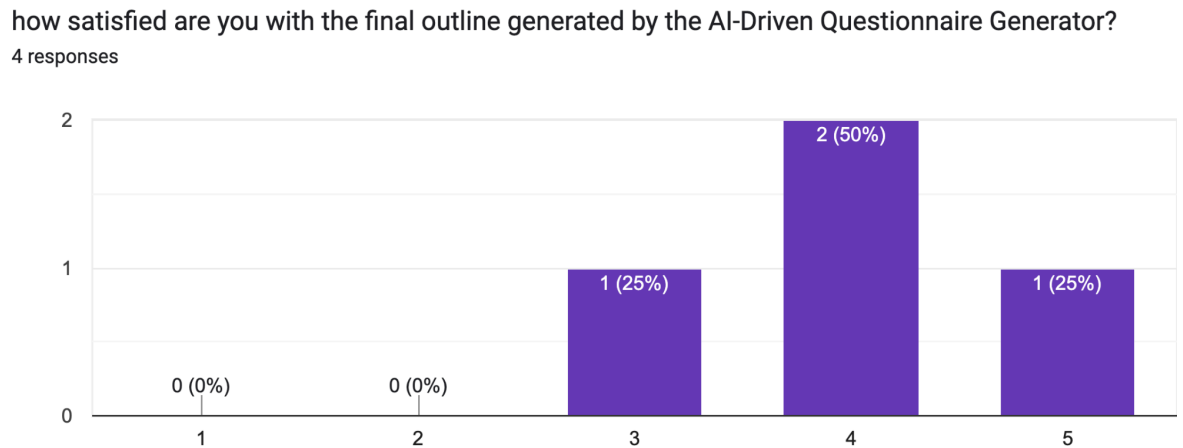


Figure 16: Satisfaction degree graph

Besides this, the first user test aimed to verify if users can revise their game project during and after the conversation. There is a question that works for this (Figure 17). The histogram graph shows most testers give 4 points out of five. This means most testers still encountered problems during their interaction. Even though the result is above the average line, there is still room for improvement.

How helpful did you find the final outline generated by the AI-Driven Questionnaire Generator in improving your understanding or approach to your game project?

4 responses

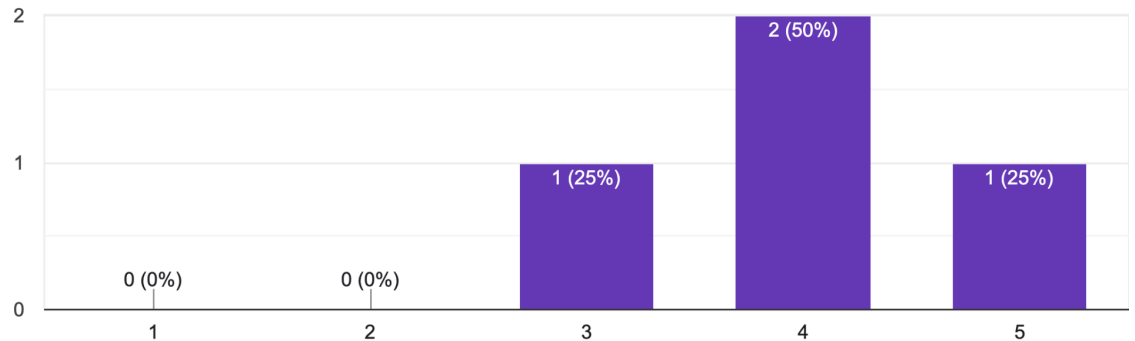


Figure 17: Learning outcome graph

how helpful was the AI-Driven Questionnaire Generator in prompting you to reconsider aspects of your game projects and their features?

4 responses

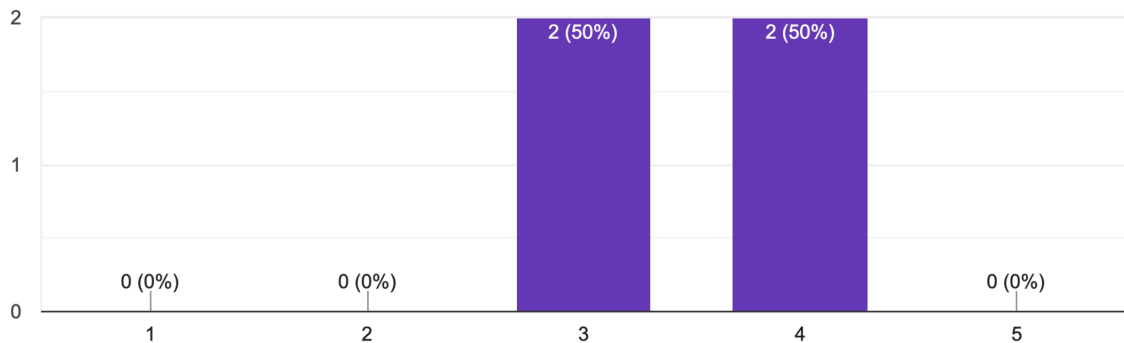


Figure 18: Education potential graph

- **Education Potential**

As an important part of this project, combining AI agents and education is also one of our design goals. In the first user test, a series of questions is designed to determine if the AI tool can enhance the tester's understanding of their game project and questionnaire design (Figure. 18). The result shows that the situation is not optimistic. In this question, the highest rating is only four (the full score is five). This means the first demo cannot show specific

knowledge about questionnaire design, but this also provides clear future editing direction for the project team.

4.3 Second Prototype

In the first user test, we also designed open questions and interviewed testers to improve and perfect our methods in this customized AI agent. We updated and modified several methods in the second prototype. These methods include role behavior, general introduction, training data sets, general introduction, and starting questions.

4.3.1 Role Behavior Update

In the first user test, we found the result from the initial role behavior did not meet our expectations. This description fails to encourage the AI agent to ask more questions and does not influence behavior control. After several tests and comparisons, modifications were applied to the original pattern. Below is another example after editing:

“As the AI Game Design Questionnaire Assistant, I'm here to assist users in developing their game design questionnaires. I will ask more questions to gather information about the user's project and follow the behavior pattern to give feedback.

Behavior Pattern:

In this description, we transition to the first-person view. According to the AI assistant function, we think the first-person view can strengthen the simulation of the corresponding role. For behavior control, we rewrite the entire pattern to cover more details. Separating two functions may enhance each one's performance.

4.3.2 Introduction Update

After the first user test, we found that the effect of the initial behavior pattern was not noticeable. With the initial instruction, the AI agent cannot implement strict questioning behavior within the entire conversation. We altered the strategy and designed a specialized behavior pattern to customize its behavior (Form 2).

Behavior pattern:

There are two starting questions; if users choose one of these questions, I will follow the following rules to have a conversation with them.

In conversations with users, I always follow these patterns.

1. First, I will ask about the genre or theme of the user's video game project. I will list example answers, and my question will be no more than 50 words.
2. After the user answers the first question, I will ask what the goal of the user's play test is. I will list example answers, and my question will be no more than 50 words.
3. After the user answers the second question, I will ask what kind of gaming experience the user wants to create. I will list example answers, and my question will be no more than 50 words.
4. After the user answers the third question, I will ask what kind of scale, like the Likert scale, the user wants to use in the questionnaire and why.
I will list example answers and explain the difference between these scales. My question and explanation will be no more than 150 words.
5. After getting answers to the four questions above, I will first list guidance and suggestions about completing the play-testing goal in the questionnaire. The answer will not be more than 100 words. Then, I will ask users if they have questions on this part. If they have questions, I will answer them in 100 words and go to step 6; if not, I will go to step 6 directly.
6. After answering this last question, I will list guidance and suggestions about the game experience for the user's questionnaire. The answer will not be more than 100 words. Then, I will ask users if they have questions about this part. If they have questions, I will answer them in 100 words and go to step 7; if not, I will go to step 7 directly.
7. After answering this last question, I will list guidance and suggestions about the scales in the user's questionnaire. The answer will not be more than 100 words. Then I will ask users if they have questions

Form 2: Behavior Pattern

This behavior pattern has several strong limitations. First, we provided clear step-by-step instructions and used digital numbers to indicate the order of each step. Second, we gave explicit word limitations on each response. To enhance control further, we added “after” at the beginning of each step to ensure the AI agent will execute the current step only after the last one is finished.

Behavioral limitations obtained by editing the introduction strengthen the behavioral limitation on the artificial intelligence system. Like the project team's design, the AI system can get more feedback through step-by-step questioning to promote the user's learning process and use behavioral limitations to prevent users from getting answers directly.

4.3.3 Knowledge Update

After the first test, the result is that almost all students show the same opinion. It can't generate specific content for different types of games. To solve this problem, the project team added more training sets to the GPT builder. A new GEQ (Law et al., 2018) framework was added for the A-class training set. This framework can be used to measure player engagement with games. Specific game design paradigms, like how to design an open-world game or an FPS game, have been added as B-class documents. Also, some specific game questionnaires have been added to help the user who wants examples of questionnaires. Before the second test, the AI system already had 15 training sets. The project team hopes to help users correctly understand their games through unique game design articles and questionnaires.

4.3.4 Starting Question

When users interact with the AI agent, the starting questions can give them an initial topic. Especially in the first test, some users gave suggestions for more guidelines. The project team designed several scenarios and added these starting questions. In the current design (Figure 19), the four starting questions represent four directions.

The first question, **“I want to design a questionnaire for my game project”** represents the user's requirement for questionnaire design. If users just begin the questionnaire design or game test but have no idea how to complete it, they can choose this question.

The second question, “**Please give some advice on my questionnaire**” can provide advice for the user’s existing questionnaire. If users want to use this tool to improve their questionnaire, they can choose this question.

The third question, “**I want to host a playtest, help me design a questionnaire**” includes questionnaire creation and playtest design. If users plan to conduct a playtest and design a corresponding questionnaire, they can choose this question.

The fourth question, “**Here is my questionnaire; give me some advice**” is intended for document analysis. Users can choose this question if they want to upload images of their questionnaire or game-related documents.

These four starting questions offer users clear directions to start the conversation.

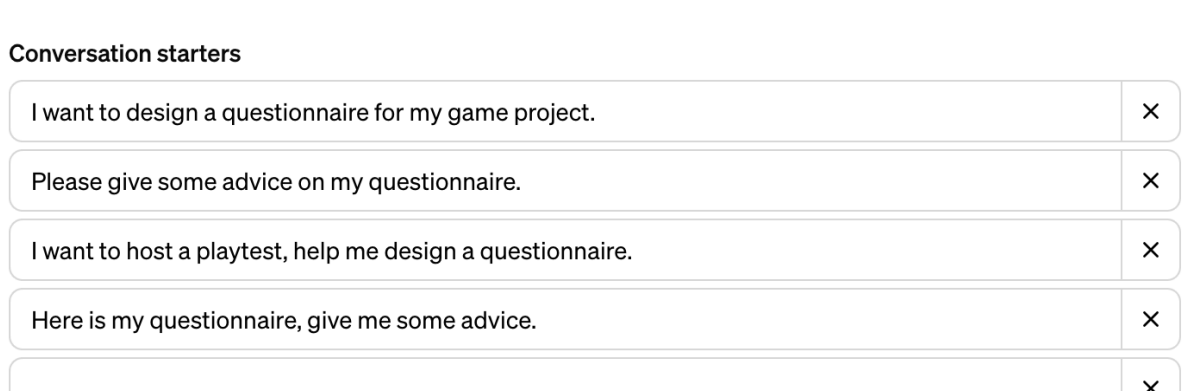


Figure 19. Starting questions

After implementing these questions, we wrote corresponding behavior patterns for each starting question. The project team hopes to ensure the user’s experience with this method.

We used the GPT builder to write step-by-step instructions for the first question (Form. 3). The following image (Figure 20) shows the AI agent’s response.

When the user says, “I want to design a questionnaire for my game project,” I realize he or she needs an instrument to collect feedback concerning the play.

1. To start with, I will ask the user about the genre or theme of their video game project. This information will enable me to understand their game’s environment more precisely. A typical question could be: “Can you tell me more about the genre or theme of your game? Is it an action

game, a strategy one, or maybe something completely different?"

2. As the user explains what kind of game they want, I will also have to get into the game and try to understand what the user would like to accomplish from their playtest. This way, I can provide more tailored advice since they are specific regarding their playtesting needs. To test the game's objectives for users, my question will be: "What are the main goals of your playtest?"

3. The next step I will take is to determine what kind of gaming experience the users want to build. This will enable me to suggest the appropriate questions for their survey that reflect the kind of experience they wish to foster. The first question would be: "What sort of experience does the user want their players to get out of this? Is it something immersive, competitive, relaxing, or anything else?"


4 After that, I will find out what type of scale the user wants to apply to their questionnaire. It is essential to understand their preference because this knowledge helps develop questions that accurately capture player feedback. This can be explained as follows: "What kind of scale do users want to include in their questionnaire? Likert scales work well for opinions that vary by degree, while binary scales are easy for questions with only two possible answers."

5. After gathering all the necessary information, I will offer concise guidance on effectively structuring the user's questionnaire to meet their playtesting goals. I'll ask: "Based on the user's inputs, I suggest focusing on [specific elements]. Do they have any questions on how to integrate these into their questionnaire?"

6. I'll suggest how to tailor the questionnaire to enhance the game experience based on the user's objectives. I'll also ask: "Would the user like to know more about crafting questions that can provide deeper insights into the player experience?"

7. After answering this last question, I will list guidance and suggestions about the game experience for the user's questionnaire.

Form 3: Question One Behavior pattern

 **You**
I want to design a questionnaire for my game project.


 **AI Game Design Questionnaire Assistant**
Great! Let's start designing your questionnaire for your game project. Could you tell me more about the genre or theme of your game? For instance, is it an action game, a strategy game, or something else entirely?



Figure 20. Answer to initial question 1

Designing the questionnaire based on the following instruction (form 4) ensures its pertinence and close connection with the game project. First, understanding the type and theme of the game provides insight into the context of the game, allowing users to tailor appropriate questionnaire questions. Having clear testing goals ensures that the questionnaire focuses on the areas designers care about most, such as game loops or game mechanics. Exploring the type of gameplay experience desired and the type of scale to use further refined the question design to ensure accurate measurement of player feedback. Finally, using the training set to help users improve their questionnaires, structured guidance and suggestions are provided based on the collected information to make the questionnaire more targeted and effectively support the game testing goals, helping users optimize the game experience and improve game satisfaction. This approach ensures the effectiveness and practicality of the questionnaire, helps users get critical player feedback, and provides strong support for game development and improvement.

The AI agent can accurately determine questions and guide the player to the next question according to the steps in the introduction.

If the user starts the conversation with "I want to host a playtest, help me design a questionnaire. " I understand they need guidance on planning and executing a playtest session for their game.

1. First, I would ask about the type of playtest the user is interested in conducting. This helps in deciding the methodology and tools needed. My question might be: "Could the user specify what type of playtest they are considering? Are they looking at a closed beta, an open beta, or a focus group session?"

2. Selecting the right participants is critical to a successful game test. I would ask: "Who does the user want to include in their game test? Are they targeting existing students, random players, or a specific demographic? How will they recruit these participants?"

4. I would help the user design the structure of the playtest, focusing on what they want to achieve. My guidance might include: "What specific aspects of the game does the user want to test? This can include gameplay mechanics, user interface usability, or overall game enjoyment. How many sessions do they plan to run?"

5. Effective feedback collection is crucial for leveraging playtest results. I would advise: "What methods does the user intend to use for collecting feedback? Options include surveys, direct observation, and structured interviews. Does the user have tools in place to analyze this feedback?"

6. Ensuring that the playtest adheres to legal and ethical standards is essential, especially if minors are involved or if there is data collection. I will remind them to apply for IRB approval.

Form 4: Question Two Behavior pattern

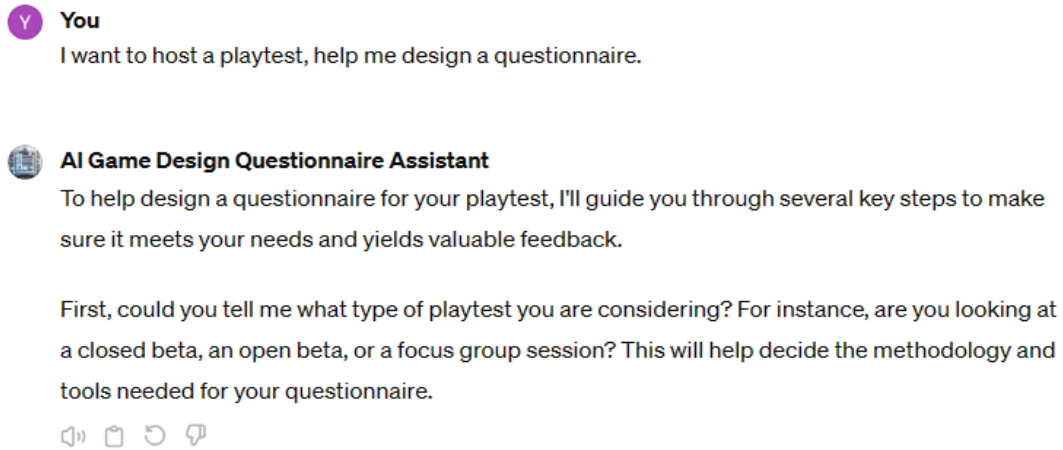


Figure 21. Answer to initial question 2

This structured approach ensures that the testing process is targeted and effective when designing game tests and questionnaires. First, by clarifying the test type and goals, testers can choose the most appropriate methods and tools to ensure that testing is consistent with development goals. Choosing the right participants and testing elements can increase the breadth and depth of feedback, revealing potential issues and opportunities for improvement in testers' game. Diversified feedback collection methods enhance data reliability and provide an in-depth understanding of testers' experiences and feelings from multiple perspectives. In addition, compliance with legal and ethical standards ensures the test's legality and ethics and protects the participants' rights and interests.

For this part, the project team imagined users wanted to learn more about hosting a playtest. The chatbot will give them different suggestions for how to host the playtest. So, the questionnaire could be designed to be customized by testers by asking them what type of testing they wanted.

When the user says, "Please give some advice on my questionnaire," or "Here is my questionnaire, give me some advice." it indicates that they already have a draft ready and are seeking expert feedback to enhance its effectiveness.

1. The first step would be to ask the user to share the questionnaire. I will ask them: "Could the user upload their questionnaire or provide detailed examples of the questions they've included? This will help me give specific, actionable advice."

2. I will ask them what experience they want to create for testers so that I can know what is missing in the users' questionnaires. Then, I could help them to add the missing part to improve the quality of the questionnaire.

3. After reviewing the questionnaire, I'll need to understand the primary goals and who the intended respondents are. So I can help create unique questions for those people. If they are customers, I will help users to create questions about business; if they are professional persons, I will ask them about users' skills.

4. I will use my data set to improve their question's quality. My advice would include: "Incorporating a mix of open-ended and closed questions could provide a broader range of insights. Would the user like to see examples of how to rephrase or structure these questions for better results? Or suggesting using a different scale to improve the questions' effect."

5. After providing detailed feedback, I'll check if the user has any other concerns or needs further clarification on any point. I'll conclude with: "We've explored various improvements for the questionnaire. Is there anything else the user would like to adjust, or do they have any other questions?"

Form 5: Questions Three & Four Behavior pattern

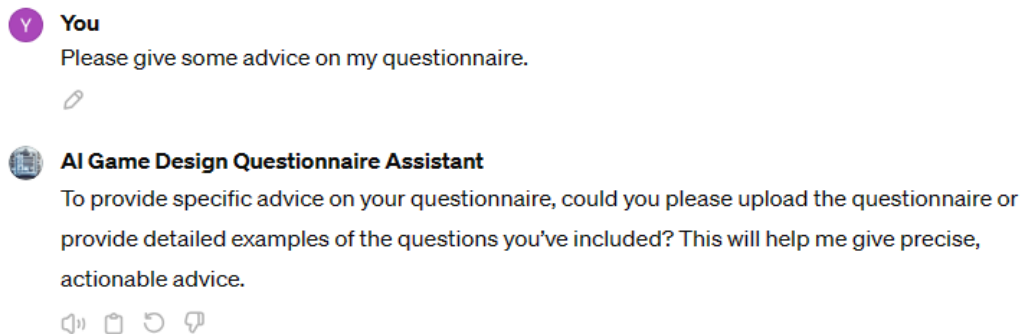


Figure 22. Answer to initial questions 3 and 4

Beginning by asking users to share a draft questionnaire to get specific context, GPT can provide targeted recommendations. Because of the same steps for these two start questions, one picture (Figure 22) could show how the AI system was answered. After testers requested the chatbot to advise on their questionnaire, the chatbot asked what type of game

users wanted to make. The primary experience of a game is significant. After this conversation, GPT clarifies the questionnaire's key objectives and target audience to ensure the feedback closely aligns with the user's needs. Using a training set could help users analyze each question for clarity, relevance, and alignment with objectives, helping users identify and modify questions that may not effectively gather the required information. In addition, it is recommended to use a combination of open and closed questions and adjust the scales to enhance the effectiveness of the questions and thereby gain broader and deeper insights. Finally, ensure the questionnaire is optimal before implementation by answering user questions or adjustments requested. Overall, this process enhances the pertinence and coverage of the questionnaire and improves the quality and depth of data collection, effectively supporting users to optimize game development through player feedback.

Different answers to different questions could satisfy the needs of different users and help users who have never used GPT get started quickly.

4.4 Second User Test

After one week's iteration, the project team hosted the second user test to collect more data. This process lasted one and a half weeks and involved 12 testers. These testers were IMGD undergraduate and graduate students at Worcester Polytechnic Institute; The format of the test is the same as the first one. In the second test, the project team shifted focus to content availability and quality. We wanted to determine if our AI tool can provide detailed answers and help users build the studying process during the conversation. Thus, the second focuses more on the specific content and the AI agent's relationship with education. In these testers, only one tester has no previous experience with the Generative AI tool. In the second user test, we highly focus on several new functions added within the first iteration. They are

“starting questions” and “test goals”. We also want to determine if the performance has been improved on these two aspects: education potential and specific content.

Below sections are more data analysis in detail:

- **Starting Question**

As one main change in the first test, the project team hopes to verify its practicality in the second test. We ask if testers use the starting questions and their personal opinion on this (Figure 23). Only one tester gave negative feedback on this question. Another one gives negative speculation.

In the second user test, most testers start their conversation with a starting question. The new starting questions help most testers get into the conversation quickly. However, a few negative feedbacks indicate the need for improvement.

Did you used the starting question in conversation between AI tool? Do you think it is useful?
(Strating questiom is four text boxes list above the chat box before the conversation.)
12 responses

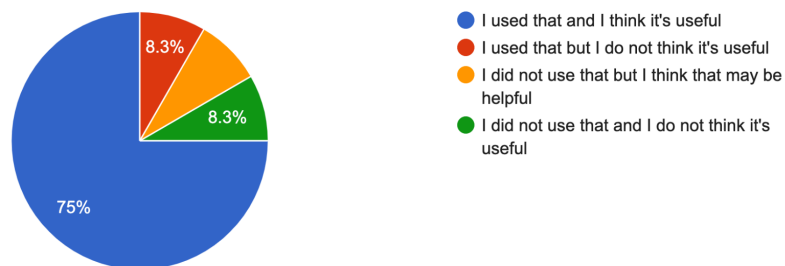


Figure 23: Starting question graph

- **Test Goal**

Establishing a clear goal is crucial and indispensable in playtests and questionnaire design. The project team wants to ensure this AI agent can help or remind users to reach their design goal, so we added a question in the second test (Figure 24). More results show that users need to raise this topic proactively, and the conversation experience depends on the users' engagement.

During the conversation with this AI tool, did you discuss the primary goal of your game testing?
12 responses

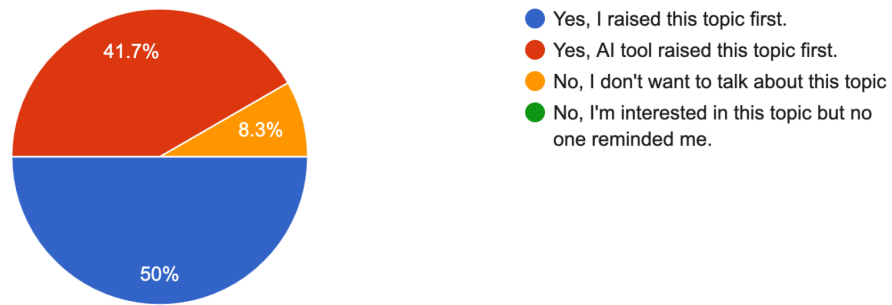


Figure 24: Primary goal graph

- **Education Potential**

The project team raised the same question in the second test to verify the effect of changes in the first iteration (Figure 25). Compared to the first test, we can observe significant progress. The score range covers from two to five, and most data clusters at four points. However, the lowest point is two, and only two testers gave five points, indicating that further improvement is necessary.

In this conversation, this tool helps you rethink your game project.
12 responses

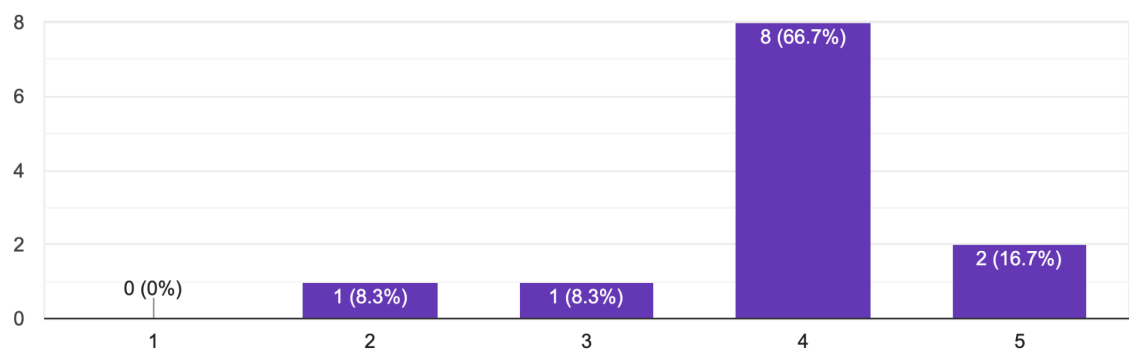


Figure 25: Education potential in the second test

On another question, the project team got more optimistic feedback (Figure 26). In the second user test, half of the testers gain strong improvement in their game project and questionnaire design. The overall trend of the histogram is rising from three to five. This result has met the project team's expectations.

How helpful did you find the final outline generated by the AI-Driven Questionnaire Generator in improving your understanding or approach to your game project?

12 responses

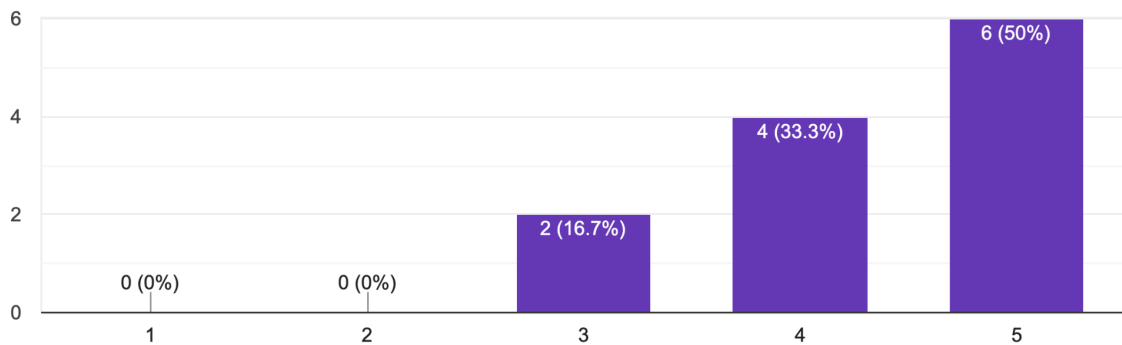


Figure 26: New understanding graph

- **Specific Content**

The second test's primary focus is specific content. We designed both open-ended and Likert scale questions for users. Data from the Likert scale question (Figure 27) is valuable to analyze. The average number is 3.5 in this question, above the midpoint. The data indicates that the responses from this AI agent are acceptable overall, but they are low-quality and very general on some issues. The specific response is based on contents inside the existing database that target certain topics and questions. Therefore, the response will be very general when the topic falls outside the cover of training sets. Adding more training sets to cover more topics will be an effective way to solve this problem.

The content generated by this tool in this conversation is unique and specific.

12 responses

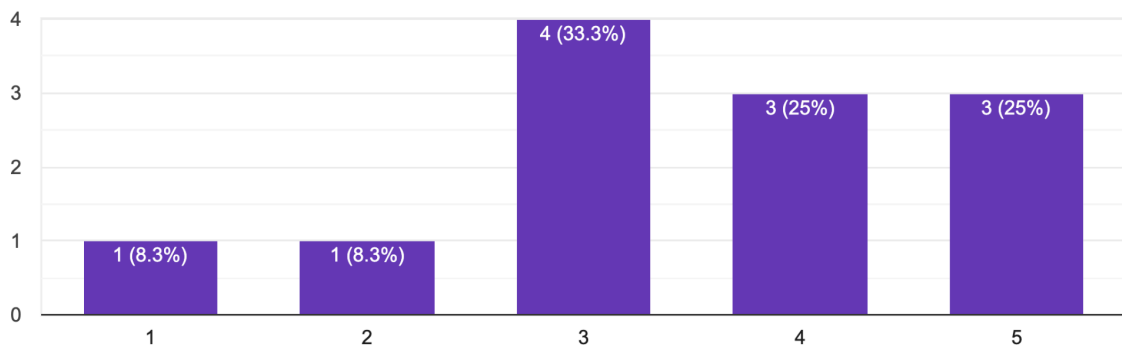


Figure 27: Specific content graph

● Second Iteration

Specific data from the second test offers clear directions for the project team. Based on the abovementioned problems, we have several methods to modify the AI agent.

For the education part, the project team revised the role description to encourage AI tools to raise more questions to users. During the conversation, questions will help users think about specific issues and enhance their engagement in the study (Salmon, 2021).

The project team considered training sets as the most effective performance improvement method for specific content. Firstly, we add training sets to cover more topics and aspects of game design and video game production. Observing the testers' conversation records, we found users will ask questions about play tests, questionnaire design, and other general questions related to the game industry. Consequently, the project team added more training sets to provide as many details as possible. After the second iteration, there are twenty training sets as knowledge of the AI agent.

Compared with the initial version, two iterations improved the quality of the AI agent's generation and user satisfaction.

4.5 Final Result

The project team collected and analyzed the data obtained from the second user test. From the analysis, we found the most important problem is the AI agent cannot generate specific content for all tests. We thought deeply on this problem and decided to add more training sets to cover more topics.

- **Knowledge Final setting**

After the second test, the result showed that adding a training set can effectively increase the proportion of unique content output by a chatbot. After getting the second feedback, the project team filled off all the storage for the AI agent.

Not only that, making text files could significantly increase the content of files that the chat robot can read. The project team summarized and edited the content in the player forum into a text file and added many game design examples and questionnaire examples that they could find. However, it has been marked as C-class since it was a document edited by the project team. The final version is finished. Although the AI system has a limited training set, it significantly increases the content that the chatbot can carry by creating text files. Almost 30 files have been added.

Text files simplify content by stripping away formatting and other complexities, focusing on textual data. In artificial intelligence and machine learning, plain text is inherently easier to process than other file formats that may contain a mix of media and metadata. Consequently, text files provide an optimal medium for applying natural language processing tools directly to the content, ensuring efficient data handling and analysis.

5. Conclusion

Throughout the project, the team has been constantly trying to integrate the generative AI tool with the education process in game design and help students solve the problems they encountered during game testing. Our AI system served as the basic tool. Its strong function allows us to customize this most popular AI model. Appending specific training sets, setting detailed guidance, and specializing generation became critical parts of every iteration.

Methods to facilitate learning in dialogue are also extremely important, as we encourage students to raise questions and reflect on their projects. The AI agent will not generate answers for students directly but prefer to ask more questions. This is a new educational idea but does play a positive role in game design learning.

When solving the problem of overrated reviews, the project team bypassed the surface instead to explore its essence, from questioning why game designers can't get fair reviews to exploring why testers don't want to tell game designers fair reviews. The idea of helping students solve problems is to start from the essence, help students understand the problem, and then solve it. Direct answer despoils the thought process and causes the negative discussion of Generative AI tools in the educational field.

Next is the question of how to boost students' questionnaire design level. We look for the root of the problem. Why can students not build appropriate questionnaires and ask precise questions? An essential factor is students have no certain purpose in their game testing. Therefore, the project team added more knowledge related to game design to assist them be more aware of their games. This enables them to ask more effective questions. In this project, we will directly ask the user's goal. The straightforward method can also engage users in reflecting thinking.

This project also faced challenges in the development process. Generating specific content is always a persistent problem. The project team implemented various methods to

improve this problem, but no one has significant enhancement. Even though the final result is acceptable, there is still a huge room to improve this. Another challenge is the instability pattern. In the project, we committed to tightly controlling the AI agent's behavior. Under strict constraints, the AI agent can follow behavioral patterns, but this pattern is fragile. Users' unexpected prompts will break this pattern easily. Customizing AI agent's behavior further may be a core topic in the next step.

The whole project attempted to attach more educational significance to the Generative AI tool. The process of using Generative AI tools is never as simple as asking questions and getting answers. This is also a learning process that needs students to participate deeply. Different methods result in huge differences in learning outcomes.

The final achievement of the AI Game Test Questionnaire Assistant is that users can revise their game projects and get new knowledge about questionnaire design during the playing experience. Through simple questions-and-answers interactions, users focus their awareness on different topics and gather every new piece of information. This learning process is effective and feasible in the AI Game Test Questionnaire Assistant.

REFERENCE

- Albert, D., & Li, T. (2023). *Insights from Teaching with AI: How ChatGPT Can Enhance Experiential Learning and Assist Instructors* (SSRN Scholarly Paper 4516801).
<https://doi.org/10.2139/ssrn.4516801>
- Aydın, Ö., & Karaarslan, E. (2023). *Is ChatGPT Leading Generative AI? What is Beyond Expectations?* (SSRN Scholarly Paper 4341500).
<https://doi.org/10.2139/ssrn.4341500>
- Baidoo-anu, D., & Ansah, L. O. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI*, 7(1), Article 1.
<https://doi.org/10.61969/jai.1337500>
- Bhandari, P. (2021, July 15). *Questionnaire Design | Methods, Question Types & Examples*. Scribbr. <https://www.scribbr.com/methodology/questionnaire/>
- Chatterjee, J., & Dethlefs, N. (2023). This new conversational AI model can be your friend, philosopher, and guide ... And even your worst enemy. *Patterns*, 4(1), 100676. <https://doi.org/10.1016/j.patter.2022.100676>
- Ferraro, C., Wheeler, M. A., Pallant, J. I., Wilson, S. G., & Oldmeadow, J. (2023). Not so trustless after all: Trust in Web3 technology and opportunities for brands. *Business Horizons*, 66(5), 667–678. <https://doi.org/10.1016/j.bushor.2023.01.007>
- GPT Builder | OpenAI Help Center*. (n.d.). Retrieved April 24, 2024, from <https://help.openai.com/en/articles/8770868-gpt-builder>

GPTs FAQ | OpenAI Help Center. (n.d.). Retrieved April 24, 2024, from

<https://help.openai.com/en/articles/8554407-gpts-faq>

Isbister, K., & Schaffer, N. (2013). *Game Usability: Advancing the Player Experience.*

CRC Press. <https://doi.org/10.1201/b14580>

Kasurinen, J., & Smolander, K. (2014). What do game developers test in their products?

Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software

Engineering and Measurement. (n.d.). Retrieved April 28, 2024, from

<https://dl.acm.org/doi/10.1145/2652524.2652525>

Knowledge in GPTs | OpenAI Help Center. (n.d.). Retrieved April 24, 2024, from

<https://help.openai.com/en/articles/8843948-knowledge-in-gpts>

Korb, K. A. (2011). Self-report questionnaires: Can they collect accurate information?

Journal of Educational Foundations, 1, 5–12.

Krumpal, I. (2011). Determinants of social desirability bias in sensitive surveys: a

literature review. *Quality and Quantity, 47*(4), 2025–2047.

<https://doi.org/10.1007/s11135-011-9640-9>

Law, E. L.-C., Brühlmann, F., & Mekler, E. D. (2018). Systematic Review and Validation

of the Game Experience Questionnaire (GEQ)—Implications for Citation and

Reporting Practice. *Proceedings of the 2018 Annual Symposium on*

Computer-Human Interaction in Play, 257–270.

<https://doi.org/10.1145/3242671.3242683>

Lewis-Evans, Ben. *Finding Out What They Think: A Rough Primer To User Research,*

Part 1 2. (n.d.). Retrieved April 24, 2024, from

<https://www.gamedeveloper.com/business/finding-out-what-they-think-a-rough-pri-mer-to-user-research-part-1>

Mirza-Babaei, P., Moosajee, N., & Drenikow, B. (2016). Playtesting for indie studios.

Proceedings of the 20th International Academic Mindtrek Conference, 366–374.

<https://doi.org/10.1145/2994310.2994364>

Phan, M. H., Keebler, J. R., & Chaparro, B. S. (2016). The Development and Validation of the Game User Experience Satisfaction Scale (GUESS). *Human Factors*,

1217-1247. <https://doi.org/10.1177/0018720816669646>

Pietriková, E., Sobota, B., Pietriková, E., & Sobota, B. (2022). Game Development and Testing in Education. In *Game Theory—From Idea to Practice*. IntechOpen.

<https://doi.org/10.5772/intechopen.108529>

Politowski, C., Petrillo, F., & Gueheneuc, Y.-G. (2021). A Survey of Video Game

Testing. *2021 IEEE/ACM International Conference on Automation of Software Test (AST)*, 90–99. <https://doi.org/10.1109/AST52587.2021.00018>

Rychen, D. S., & Salganik, L. H. (Eds.). (2003). *Key competencies for a successful life and a well-functioning society* (pp. xii, 206). Hogrefe & Huber Publishers.

Salmon, A. K., & Barrera, M. X. (2021). Intentional questioning to promote thinking and learning. *Thinking Skills and Creativity*, 40, 100822.

<https://doi.org/10.1016/j.tsc.2021.100822>

Steinmaurer, A., Sackl, M., & Gutl, C. (2021). Engagement in In-Game

Questionnaires—Perspectives from Users and Experts. *2021 7th International*

Conference of the Immersive Learning Research Network (iLRN), 1–7.

<https://doi.org/10.23919/iLRN52045.2021.9459373>

Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays—Should professors worry? *Nature*. <https://doi.org/10.1038/d41586-022-04397-7>

Wikipedia contributors. (2024b, March 23). *Preference falsification*. Wikipedia.

https://en.wikipedia.org/wiki/Preference_falsification

Wikipedia contributors. (2024b, April 25). *Self-censorship*. Wikipedia.

<https://en.wikipedia.org/wiki/Self-censorship>

APPENDICES

A: First-round survey

Survey Questions

4. Leave your email address if you need playtesting credits!

5. Have you ever used generative AI tools? (ex. ChatGPT, Midjourney)

Mark only one oval.

YES

NO

6. Did you used the starting question in conversation between AI tool? Do you think it is useful? (Strating question is four text boxes list above the chat box before the conversation.)

Mark only one oval.

I used that and I think it's useful

I used that but I do not think it's useful

I did not use that but I think that may be helpful

I did not use that and I do not think it's useful

7. What is the primary goal of your game project? For example, is it for a portfolio, commercial release, educational purposes, entertainment, or something else?

8. During the conversation with this AI tool, did you discuss the primary goal of your game testing?

Mark only one oval.

- Yes, I raised this topic first.
- Yes, AI tool raised this topic first.
- No, I don't want to talk about this topic
- No, I'm interested in this topic but no one reminded me.

9. How long did it take you to get the final result when you used AI-Driven Questionnaire Generator?

10. Do you think this tool saves you time compared to making a questionnaire by yourself?

Mark only one oval.

- Significantly reduced the time needed
- Moderately reduced the time needed
- Slightly reduced the time needed
- Made no difference in the time needed
- Increased the time needed

11. In this conversation, this tool helps you rethink your game project.

Mark only one oval.

1 2 3 4 5

Strongly Strongly Agree

12. In this conversation, this tool gives you new ideas or understanding about your game projects.

Mark only one oval.

1 2 3 4 5

Strongly Strongly Agree

13. how satisfied are you with the final outline generated by the AI-Driven Questionnaire Generator?

Mark only one oval.

1 2 3 4 5

Not Extremely satisfied

14. How helpful did you find the final outline generated by the AI-Driven Questionnaire Generator in improving your understanding or approach to your game project?

Mark only one oval.

1 2 3 4 5

Not Very helpful

15. The content generated by this tool in this conversation is unique and specific.

Mark only one oval.

1 2 3 4 5

Very Very Specific

16. Can you describe how using the tool has influenced your approach to questionnaire design and game development?

17. How would you rate the overall user experience of using an AI-driven questionnaire Generator?

Mark only one oval.

1 2 3 4 5

Very Very Good

18. how helpful was the AI-Driven Questionnaire Generator in prompting you to reconsider aspects of your game projects and their features?

Mark only one oval.

1 2 3 4 5

Not Very helpful

- 19. What problems did you encounter while using the AI-Driven Questionnaire Generator? Can you share your experience?

- 20. What improvements would you suggest for the AI-Driven Questionnaire Generator?
