Universiteit
Leiden
The Netherlands

Bachelor Computer Science

Teaching the Importance of Critically Reviewing
the Output of Generative AI through Games

Alette Farzad

Supervisors:
G. Barbero & Dr. A.N. van der Meulen

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

08/08/2025

Abstract

There has been a notable rise in the use of Generative Artificial Intelligence; however, the level of AI literacy has not kept pace. It is thus essential that we teach end-users the significance of digital skills like output review. In this regard, educational games are a compelling option because of their ability to foster intrinsic motivation and to promote the development of computational thinking skills. Moreover, educational games promote the development of computational thinking, a crucial skill for cultivating AI literacy.

We aim to investigate students' acceptance of educational games as a tool for learning about AI literacy by analyzing their behavioral intentions using the Technology Acceptance Model. Additionally, we investigate the design process required for the development of these educational games. We developed an educational game called Doolhof in which players explore a maze and solve puzzles in search of treasure. Players are provided with a handbook and a 'Robot Companion' to assist them. This companion is presented as generative AI and may occasionally produce inaccurate or irrelevant responses to encourage players to critically assess the output by using the provided handbook.

An empirical study involving 17 students (ages 16-17) was conducted in which participants played Doolhof for 25 minutes. Subsequently, they answered a questionnaire consisting of three sets of questions: direct questions on a Likert scale, questions presenting hypothetical scenarios for participants to navigate, and open-ended reflective questions. Participants reported favoring educational games such as Doolhof over traditional lesson methods. Furthermore, they reported understanding the importance of output review. Despite this, only half of all participants performed output review during the hypothetical scenarios.

We conclude that educational games have the potential to form an effective tool for teaching AI literacy due to their engaging nature. Results indicate Doolhof was generally favored over traditional lesson methods and was effective at raising awareness of the importance of output review. However, Doolhof had limited success in further motivating students to perform output review. Future research could explore and optimize the design process of educational games in the context of AI literacy.

Contents

1 Introduction

In the last decade, a trend of introducing game elements in education has emerged. Game elements in educational contexts are found to have several benefits, such as promoting students' engagement and inviting intrinsic learning motivation through their design [62]. The unique combination of challenges and feedback that game elements offer make them able to teach complex topics in a stimulating way [28]. Furthermore, educational games are found to have the ability to stimulate Computational Thinking [52], a set of problem-solving skills encompassing decomposing problems into smaller parts: pattern recognition, abstraction, and algorithmic thinking [59]. These skills form an important basis for addressing modern problems, enhancing individuals' critical thinking, creativity and adaptability.

Moreover, with Artificial Intelligence (AI) technologies becoming more prominent, games present themselves as a unique ground of interaction between humans and AI. As such, games specifically designed for education could leverage this ground of interaction for educating students on the use and nature of AI [18]. Especially now, as students grow to use AI tools more and more, ensuring that they know how to use these technologies properly and responsibly is critical [6]. Recent developments within education aim to address this educational gap, focusing on fostering digital skills [47]. Educational games fit nicely within these efforts, supporting skills such as computational thinking that are also necessary to foster digital literacy, as well as having the potential to ease adoption of these new educational frameworks within the classroom. As such, educational games are a promising tool to help equip students with the knowledge to navigate and thrive in the modern digital landscape.

1.1 Problem Statement

The use of Artificial Intelligence (AI) has become widespread in recent years. Specifically, the use of generative AI (genAI) has increased, especially among younger generations. In a report written in collaboration by Common Sense Media and Hopelab, it is reported that about 27% of ages 14-22 occasionally use AI, with about 4% of the total sample using generative AI almost daily [15]. This surge in usage can be partially attributed to how easily accessible generative AI such as ChatGPT [1] have become. With a simple query, one can ask for help in summarizing lengthy papers, finding cooking recipes, or even generating fragments of code. Generative AI proves itself as a useful tool that many have started to rely on greatly [50].

However, generative AI has certain limitations. Due to their stochastic nature, generative AI is prone to outputting answers that, while appearing realistic, turn out to be incorrect or nonsensical with the given prompt [58]. OpenAI acknowledges this limitation of ChatGPT [43], stating that fixing the issue is challenging due to the inherent limitations of the model design. This issue is not unique to ChatGPT, as it is suggested even the most modern and advanced generative AI holds the same issues [64]. With this in mind, it becomes clear that blindly relying on the output of generative AI would be ill-advised. The younger generation in particular forms a vulnerable group [46]. For example, incorrect or inappropriate information given by generative AI related to mental health issues could enable dangerous behavior [38]. In addition, the common use of generative AI by students for their schoolwork means that unless students treat generative AI responsibly, their study

[1] ChatGPT: https://openai.com/index/chatgpt/

performances can be negatively impacted [15]. Therefore, it is vital for the students to learn about the stochastic nature of AI at an early age and to critically assess generated output. We define this skill as output review: the skill of verifying the output of generative AI by cross-referencing sources and information. We more formally define output review and elaborate on why we choose it as a starting point in Section 2.4.1.

1.2 Research Questions

This thesis aims to research how educational games can be used to promote output review within the context of generative AI. We propose the following research questions:

Research Question
How can games be used to teach the importance of critically reviewing the output of generative AI?

> *Sub-question 1*
> Why is it important that we teach the importance of critically reviewing the output of generative AI?
>
> *Sub-question 2*
> How does one design a video game that motivates players to critically review the output of generative AI?

As a part of this thesis, an educational game and its corresponding design documents are to be delivered.

1.3 Overview

This work is a bachelor thesis project at the Leiden Institute of Advanced Computer Science (LIACS) and supervised by G. Barbero (LIACS) and Dr. A.N. van der Meulen (LIACS). This chapter served as the introduction; Section 2 outlines the background and related work and introduces relevant terminology; Section 3 discusses methodology for the project, including the game design process and the empirical study; Section 4 presents the final product of our game design process and the results of our empirical study; Section 5 serves to discuss our results and limitations. Finally, in section 6 we answer our research questions and describe future work. Acknowledgments regarding this thesis are given in section 7. Our appendices contain relevant game design documents expanding on the characteristics (Appendix A) of our game and detailing its original design document (Appendix B), as well as other graphs of data that were not directly included in the results section (Appendix C).

2 Related Work

In this section we give an overview of the related work for this thesis. We introduce terminology regarding games in education in section 2.1. Subsequently, in section 2.2 we discuss game design from an educational perspective and introduce some theories and frameworks that we wish to use for this goal. Section 2.3 discusses the current state of generative AI and its limitations to further expand on the need for teaching responsible AI usage. Section 2.4 goes over current efforts within both the Dutch education landscape and the global progress regarding AI education. Additionally, we define the concept of output review in 2.4.1 and discuss the relevance of computational thinking skills for responsible AI usage in 2.4.2. Finally, section 2.5 introduces a framework to measure user acceptance of technology. We wish to use this framework to investigate students' acceptance as a tool for learning about AI literacy.

2.1 Game Elements in Education

With game elements becoming increasingly more common within educational contexts, various approaches to applying game elements have emerged. Among these approaches, a distinction can be made between Gamification and Game-Based Learning. To begin with, Gamification is defined as the use of game design elements within non-game contexts [11]. Game design elements here refer to common concepts used within games, such as leaderboards, point scoring, and streaks that gamification leverages to enhance competition and motivation. In order for gamification to be effective, each game element is linked to a specific learning outcome [29]. This transforms the typical learning experience to feel more like a game. The main design, however, still revolves around the learning goal. Gamification has positive effects when used in an educational context, having been found to boost students' learning motivation [34].

	Game for Learning (G4L)	Game-Based Learning (GBL)	Gamification
Basic Definition	A game ***designed*** specifically with some learning goals in mind.	The process and practise of ***learning*** using games. [From the **learner's** point of view]	The use of game elements in a non-game context.
Purpose	Normally connected with some educational goals.	**Not a game** - this is an approach to learning.	Often used to drive motivation, but can also be used to make something more playful and game-like.
Primary Driver (Why used?)	To learn something.	To improve learning. To improve learning effectiveness.	Depending on how it's implemented, it can tap into extrinsic or intrinsic rewards (or both)
Key Question	Is it effective?	Am I learning what I am supposed to be learning?	Education: Is it effective?
Focus	Content / Message (what)	Learning Objectives (what & how)	User Experience (how)
Concept Catalyst	Performance or Knowledge Gap	Game is the lesson or is used as part of the lesson	In learning it usually impacts HOW things are taught and administred rather than WHAT is taught.
Fidelity	Faithfulness to message essential	Faithfulness to message essential.	Not applicable. If a narrative exists, it need have nothing to do with what's being gamified.

Figure 1: Excerpt of K. Becker's *What's the difference between gamification, serious games, educational games, and game-based learning?*[3]

In contrast, Game-Based Learning (GBL) is an approach to learning where the learning happens through playing games. The value of GBL is more intrinsic, embedding learning holistically within the game itself to make it enjoyable and engaging. GBL is a valuable pedagogical approach that has been found to stimulate students' motivation and also enhance their joy, autonomy, critical thinking, creativity, and imagination [28]. In addition, it can be applied well within both online and physical environments, making it especially suited to meet modern educational requirements. This pedagogical approach often leverages educational games. An educational game, otherwise referred to as a Game for Learning (G4L) is a game specifically designed with some learning goals in mind. The embedding of the learning content is intrinsic and intentional in the design, rather than it being a byproduct. Figure 1 outlines the difference between gamification, G4L and GBL in further detail.

There already exists a precedent of utilizing games for AI education. The majority of these games revolve around coding. However, some of these aim to educate children of ethical and social implications and work on explaining the core concepts of responsible AI usage [14]. Additionally, studies have been conducted regarding the effectiveness of GBL and educational games for fostering AI competencies and skills [19, 12]. These studies highlight how GBL effectively enhances students' comprehension of AI, particularly when integrated with problem-based learning, as this allows for further development of students' critical thinking and problem-solving skills [19]. Educational games are therefore suitable tools for teaching about responsible AI usage.

2.2 Game Design for Education

Game-Based Learning has a lot of potential in furthering competencies necessary for responsible AI usage. With the design process of educational games being more focused on attaining specific learning goals, it becomes worthwhile to look at ways in which we can structure the game design process to ensure these learning goals are met. Various theories exist that, when used as guidelines, can aid us in the design process [16]. In this section, we discuss three theories that originated outside of a game design context but have been found to be effective guidelines for game design. Lastly, in section 2.2.4 we discuss the main framework with which we intend to develop our educational game.

2.2.1 Bloom's Taxonomy

Bloom's Taxonomy is a hierarchical framework designed to classify learning objectives that originated in the context of education. It was originally proposed in 1956 by Benjamin Bloom and was more recently revised by Krathwohl and Anderson [26]. This framework outlines the different categories of knowledge, structured from simple to more complex and challenging types of thinking: remembering, understanding, applying, analyzing, evaluating, and creating. Figure 2 shows how the framework is ordered. By framing learning goals using these verbs, we can define learning goals that progressively advance students' skills. There is merit to applying this framework to game design, as especially within the design of educational games, it is easy to lose sight of what we wish to teach. Clearly outlining the learning objectives allows us to design educational games more progressively, ensuring that the embedded content remains effective at teaching. This way, we can use Bloom's Taxonomy as a basis for creating our game design [53].

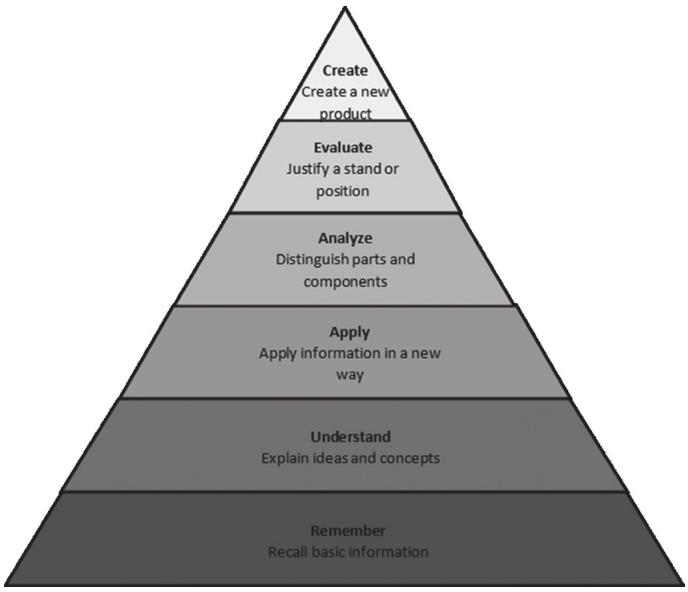


Figure 2: The Revised Bloom's Taxonomy of Educational Objectives. From Jensen et al. *Beyond Bloom's: Students' Perception of Bloom's Taxonomy and its Convolution with Cognitive Load* [21]

2.2.2 Goal-Setting Theory

Goal-setting theory, designed by Edwin A. Locke and Gary P. Latham [32], specifically examines how defining goals can improve performance. In educational games, vague goals and objectives may cause frustration, which can disrupt the state of flow. Furthermore, they may also distract the player from the main objective. Applying Goal-Setting theory can thus greatly benefit G4L design. Goal-setting Theory has five main principles, which we apply to a game-specific context below.

Firstly, objectives within the game should be specific. It should be clear what the player is meant to do. Goals also need to be engaging, inviting commitment. Ideally, a player needs to be invested in the outcome - for example, by expecting a reward for achieving a goal. Goals should also be tough but not feel impossible to achieve. This principle of inviting challenge ties in with Flow theory. Furthermore, a player should receive feedback for progressing through goals. Lastly, goals should not be too complex. Ideally, complex goals are broken down into smaller sub-goals. Utilizing these principles, we are able to set well-defined goals, which benefits player motivation and performance [16].

2.2.3 Flow Theory

Flow Theory was founded by Csikszentmihalyi [8] and is used to describe a state of optimal experience that lies between challenge and skills. In this state, a person is fully immersed in a feeling of focus, involvement, and enjoyment in the process of an activity. This state can be achieved within educational games. During play, too little challenge leads to boredom, while too much may lead to anxiety. Balancing the activity's challenge level with the player's skill level properly is a big task, but one that yields a situation in which the activity is engaging and stimulating.

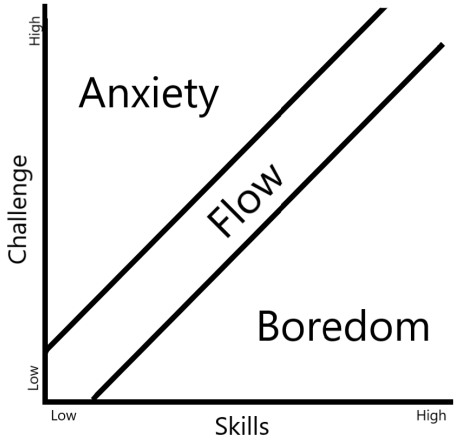


Figure 3: Diagram of Flow Theory. After Mihaly Csikzentmihalyi, *The Flow (1990)*, p.74

This is vital for educational games: We want as much engagement as possible, as this is when the learning process is enhanced best [33]. Figure 3 illustrates this balance. For game design, this means that an iterative design process is vital: Challenges designed within the game environment need to be tested to ensure they are difficult enough while at the same time not frustrating the player.

2.2.4 DFV Framework

The theories we just discussed are useful guidelines when it comes to improving educational games, providing us with ways we can balance challenge and skill, improve challenges, and identify the learning goals we wish to embed. However, when working from a blank slate, these theories provide little structure in conceptualizing a basic design and are instead used to iterate on already existing concepts. An unstructured approach to educational game design has the potential to cause the educational content to be embedded poorly. This means that we must take a wider view than merely adhering to these guidelines and put our design process itself under scrutiny.

For the purposes of this thesis, we utilize the Desirability, Feasibility, and Viability framework [20]. This framework was popularized by IDEO for product design. Deconstructing our product to these essential values allows us to craft the core requirements for the design of our game, allowing for a proper starting point that can be iterated over to arrive at a final concept. We used a modified version of this framework to allow us to ensure our game design fulfills both user needs and achieves the defined education goals. The first of the framework items, Desirability, brings into perspective user appeal, asking how we can fit the game to user needs best. The next item, Feasibility, asks if this educational game is operationally possible to make. What are the technical requirements and constraints it must adhere to? Lastly, Viability concerns the practical requirements within the game to ensure the core learning values are well-embedded. In the original model, the latter focuses on financial viability, but for the purposes of an educational game with no development budget, this is redundant, and we require a method to adhere to our learning goals. With this in mind, we are able to identify essential points that ensure our design is feasible, viable, and desirable. We go further in-depth on how we achieve this in section 3.

2.3 State of generative AI

To properly design a game regarding critically assessing the output of generative AI, we must first take an in-depth look at the state of the art of genAI, both in a wider context and within that of education. We do so in this section. Additionally, we wish to identify the limitations of this current state of generative AI, which we discuss in section 2.3.1. Lastly, we discuss the current state of education in regard to responsible AI usage in section 2.4.

Generative Artificial Intelligence is a term that refers to a collection of various technologies that produce text, images, video, or other forms of data through predictions based on patterns learned from large amounts of existing data. Their predictive nature makes them well-suited for various applications, such as the so-called Large Language Models (LLMs), which are a subset of generative AI that is trained on vast amounts of text in order to be used for natural language processing tasks. In recent years, many LLM applications of generative AI have appeared, such as ChatGPT [2], Google Gemini[3] and DeepSeek [4]. These are commonly used for answering search queries, text generation, and language translation. Additionally, LLMs exist that are particularly tuned to helping with specific tasks, such as generating pieces of code based on user prompts. Applications of generative AI also go beyond LLMs and beyond daily personal use, nowadays seeing applications being tested in sectors such as business and healthcare [39].

With the advent of this rapidly developing technology, generative AI has also started appearing within educational contexts. With these technologies, learning content can be transformed to students' needs. LLMs can be applied by making summaries or flashcards of learning content. Furthermore, as using ChatGPT for schoolwork illustrates, generative AI can provide students with writing support and more research capabilities through enabling better search queries [15].

Various studies show potential beneficial effects of using AI in education. A study from 2023 by Michael Sailer et al. performed an experiment with AI-generated automatic feedback for pre-service teachers [48]. This experiment found that, when compared to static feedback, in simulation adaptive feedback can give scalable and process-oriented feedback in real time to many students in higher education. Furthermore, experiments have been done with AI tutors in which students were shown appreciating the AI tutor's immediate and personalized support [2]. While students primarily saw it as complementing human tutors, rather than replacing them, it provided them with a space where they could ask questions without being judged by others. In other fields, it has been shown that using AI within education could boost student motivation and engagement [37].

It is clear that in the modern day these rapidly evolving AI technologies offer unique ways to innovate various fields such as education. As highlighted with the aforementioned studies, different applications already exist. That said, there is still plenty of room for improvement and research - both in developing the AI systems themselves and in further researching the impact of using AI technologies in these fields. Moreover, there are also limitations to be considered within these technologies that further impact their use, which we discuss in the following section.

[2]ChatGPT: https://chatgpt.com/
[3]Gemini: https://gemini.google.com/
[4]DeepSeek: https://www.deepseek.com/

2.3.1 Limitations of generative AI

Despite rapid advancements, generative AI is not without limitations. As discussed earlier, these technologies are prone to producing output that, despite seeming realistic, is nonsensical, improbable, or otherwise incorrect. This phenomenon is referred to as "hallucinating". We distinguish three types of hallucinations that generative AI can produce [63]. The first of these types is Input-Conflicting hallucinations, where a response is given that contradicts the user's input. For example, this would occur when you ask an LLM what to wear on a sunny day, but the LLM responds with what to wear on a rainy day. These types of hallucinations typically occur when there is a misunderstanding by the generative AI of user intent. Secondly, Context-Conflicting hallucinations occur when the genAI either ignores or adds things to the context of a situation. Regarding our previous LLM scenario, this would occur when the LLM adds in an extra person that was not mentioned in the original prompt. This can occur due to limitations in maintaining long-term memory or identifying relevant context and is thus a structural problem. Finally, Fact-Conflicting hallucinations are simply the AI producing something that is factually untrue, such as an LLM claiming that the earth is flat. There are multiple sources of Fact-Conflicting hallucinations, but incorrect training data is a common cause.

Understanding the nature of hallucinations as different types of false information allows us to better design a G4L that aims to convey why one should be aware of the potential for AI to hallucinate. For the purposes of this research, our main focus will be on fact-conflicting hallucinations, as those often go undetected when the user lacks knowledge of said facts and often have the greatest impact. In contrast, input-conflicting and context-conflicting hallucinations are more easily noticed. Additionally, it is vital to note that hallucinations are an integral challenge within the design of generative AI systems. This means that the problem of hallucinations will likely remain relevant for a long time, as even the most modern generative AI currently hallucinate [64]. Furthermore, there are other limitations to generative artificial intelligence beyond hallucinations [63]. While our main focus in this thesis will be on making users aware of false-information hallucinations and the need to check the output for correctness, it should be kept in mind that answers generative AI gives may also suffer from ambiguity, incompleteness, under-informativeness, or bias. Furthermore, for responsible AI usage, one ought to be critical with the type of tasks that they use generative AI for. After all, even the best LLMs have poor reasoning capabilities [36].

Especially within educational contexts, further concerns are present regarding the use of AI technologies. Educators report they face significant challenges with the adoption of AI within the classroom, the largest among these being students using AI for cheating [30]. Additionally, the use of generative AI also heightens the risk of accidental plagiarism. The nature of genAI, being a system trained through digesting large amounts of previously published work in order to create output, causes it to occasionally reproduce existing content. Without critically assessing the output, this could lead to accidental plagiarism being committed. Perhaps a more pressing concern is whether AI systems actually support learning: Recent studies have shown varying results regarding the impact of AI on learning. Frequent AI usage has been shown to not necessarily translate into better academic outcomes [31]. Using generative AI for certain skills such as programming is not a functional substitute for teaching these skills either, and over-reliance might instead hinder learning [49]. Finally, a recent study by MIT's Media Lab suggests that over-relying on generative AI could actually harm learning, especially for younger users [25]. While further research still needs to

be done to explore the full depths of these limitations and their effects on learning, it becomes abundantly clear that despite the many advancements within the field of AI, responsible AI usage is not a skill we can neglect. Additionally, educators report challenges in the adoption of AI systems from a lack of training or support, which hinders the integration of AI in education [30]. This highlights that there is interest in tools that could aid teachers in the advancement of AI education, such as educational games.

2.4 AI Literacy

In the past year, developments have been made to digital literacy requirements within the Netherlands. Digital literacy in this context refers to the competencies necessary to participate in the modern digitalized society. These competencies now fall under the basic skills that elementary and secondary schools should aim to teach [47]. Stichting Leerplan Ontwikkeling (SLO)[5] has been tasked with researching how digital literacy can be best integrated within the lesson curriculum. Digital literacy is defined over four domains that encompass skills such as searching for information about digital technology, the use of digital technology, the critical use of said technology, and the ability to estimate risks of the use of digital technology [42].

The SLO has outlined six core goals that together form a framework to encompass teaching digital literacy [23]. These core goals can be used by educators to shape their lesson plan for teaching skills and competencies such as teaching data safety, being able to navigate around false information online, and understanding the importance of digital privacy and safety. Specifically, a separate core goal has been designed for Artificial Intelligence competencies. There were mixed opinions of the adoption of this core goal as a separate point within this framework, but ultimately AI was deemed as an integral part of the digital landscape in modern times, as many commonly used tools and software already integrate AI [23]. This core goal primarily encourages students to be taught to explore the possibilities of AI systems, being able to both recognize and use AI systems and tools in their surroundings. An aspect of this not explicitly mentioned within this core goal of AI is the need to treat AI critically due to the risks that the aforementioned limitations pose, despite the critical usage of technology being defined as a domain of digital literacy [42]. Applying this definition to this core goal, we thus assess that it is necessary to not only teach how to effectively use AI tools but also to do so responsibly in order to alleviate these risks.

To address this need for responsible usage, the term "AI literacy" has emerged, referring to all the competencies and knowledge necessary for critically engaging with and using AI technologies [40]. With the potential risks of AI, even the Dutch government has mandated that those working with AI have a moderate amount of AI literacy [44]. Examples of skills necessary for AI literacy include a basic understanding of how AI works, being able to recognize it in practice, understanding the limitations and biases of AI technologies, and understanding the ethical considerations and societal impacts of AI systems. UNESCO has developed an AI competency framework for students that outlines this in further detail [35]. The UNESCO framework expands thoroughly on what the Dutch literacy core goal of AI touches on, providing an in-depth definition of AI literacy.

[5]Stichting Leerplan Ontwikkeling: https://www.slo.nl/

2.4.1 Output Review

An important skill not explicitly defined within the UNESCO framework is that of Output Review. We see that point 4.1.1 in table 2 states that "Students will understand what it means for AI to be human-controlled, and what the consequences could be when that is not the case" [35]. Furthermore, the associated curricular goals state that AI curricula should "facilitate an understanding on the necessity of exercising sufficient human control over AI". While this implies human verification over generative Artificial Intelligence is necessary rather than blindly utilizing the output, there is some vagueness as to what extent human control is necessary. We wish to further extend this framework by formally defining "Output Review" as follows. A user performs output review when they aim to critically analyze generative AI output, intending to identify any hallucinations, ambiguity, bias, incompleteness, or under-informativeness that might be present in the output. When using this output, the user then understands the implications of these factors present and either modifies the output, aiming to remove these factors, or uses the output with these factors in mind.

Proper output review is difficult, as in many cases it might be impossible to detect problem factors with generative AI output, especially as modern genAI become more advanced. Despite this difficulty, it remains important to critically assess output, as even if not all problem factors can be caught, performing output review still mitigates the risk they pose. With there being no guarantee that output review actually catches all errors, we wish to define output review by the **intention** to treat the output critically, rather than necessitating the full elimination of such factors. As a practical example, when a student uses an LLM for retrieving information for an essay, it is the expectation that they analyze the output and attempt to cross-reference any information that the LLM gives them with other sources. In this case, if despite this attempt they fail to identify a piece of incorrect information, they would still have performed output review.

Critically assessing the output of generative AI is relatively non-complex as a concept, yet is a rather crucial aspect of mitigating risks when using AI regularly. Even if output review is not always fully effective with catching errors, it still promotes critical thinking regarding the prompted subject, lessening the aforementioned risks of over-reliance on generative AI. In education specifically, this means engaging students in the learning process and encouraging them to treat the materials more critically, mitigating the effects of over-reliance on learning. While AI literacy efforts currently primarily focus on fostering a basic understanding of AI concepts and responsible usage, they lack sufficient detail in instructing the practical application of these concepts. Therefore, extending these frameworks through thoroughly outlining concepts such as output review allows us to further facilitate the practical application of responsible AI usage. With output review being relatively non-complex as a concept yet rather effective at alleviating pitfalls, output review forms a great focus point for furthering AI literacy after an initial understanding of AI technologies has been achieved. Finally, it should be noted that the use of output review remains fairly heavily influenced by user motivation. Even if a user were to be aware that AI might hallucinate, they might not always take the effort to do an exhaustive review of the output. Considering this limitation, educational games form a suitable method for encouraging output review due to their ability to instill intrinsic motivation [62].

2.4.2 Computational Thinking

Computational Thinking (CT)[59] skills are a fundamental part of digital literacy, being defined as a separate domain by the SLO [42]. These skills also form a valuable basis for teaching AI literacy. For example, understanding algorithmic thinking directly translates to relevance when it comes to designing, developing, and using AI tools and applications. In general, computational thinking skills contribute greatly to critical thinking skills [27], which in turn contribute to the understanding of AI systems. This understanding then contributes to understanding the relevance and importance of output review, serving as motivation to do so. Furthermore, it has been theorized that there is a link between CT and Prompt Engineering [13], which is the act of crafting and refining inputs to AI models to generate desired outputs. This falls under AI literacy. CT is required for Prompt Engineering, while at the same time, a good grasp of Prompt Engineering enhances the learning of Computational Thinking skills [13]. This is part of a wider set of research that states that AI education benefits the development of computational thinking skills [60, 24]. However, while various research indicates a link between the two [56], the exact nature of how Computational Thinking skills can benefit AI literacy remains underexplored. That said, it can be inferred that as with Prompt Engineering [13], computational thinking helps with decomposing and analyzing the concepts introduced within the wider scope of AI literacy. Furthermore, it might further enhance the structural thinking and critical thinking required for output review. Therefore, CT skills help build towards stronger output review skills.

2.5 The Technology Acceptance Model

The Technology Acceptance Model [9, 10] is a way of predicting user acceptance of software. It identifies two main factors that it claims significantly influence a user's attitude towards a technology, which in turn affects their intention to use it and ultimately determines their actual usage behavior. These two input variables are Perceived Usefulness, which refers to how a user believes that the technology will improve their performance or help them achieve their goals. Perceived Ease of Use, the other factor, refers to the degree to which a user believes the system will be convenient and effortless to use. The Attitude towards usage reflects overall feelings of the user towards the product. Behavioral Intention to use is an outcome variable that measures the user's intention to actually use the technology [10], thus being an important factor that we wish to optimize as it is what ultimately leads to actual system use [54].

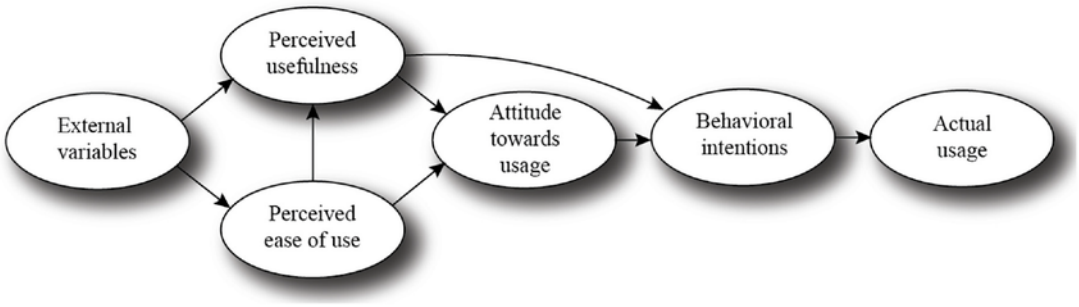


Figure 4: Extended version of the Technology Acceptance Model. From Haverila et al. [17]

We wish to utilize the TAM model to determine players' acceptance of our designed game as a method to teach about output review, as well as their general disposition towards utilizing educational games for AI literacy education. The TAM model has been used prior in similar contexts, yielding significant results. A study by Manuel Ninaus et al. investigated the acceptance of GBL by students. Additionally, they aimed to investigate which individual variables could lead to learning success within GBL [41]. Their findings suggest that learning success is predicted by the perceived usefulness of the game as a learning tool, as well as the ease of use. In addition, it was found that students' learning success was also influenced by their intrinsic motivation for the respective learning domain. The Technology Acceptance Model is well-suited to measure these factors. Furthermore, the TAM model has also been used to investigate the acceptance of GBL by educators, where it was found that a significant factor of the acceptance of GBL by educators was the quality of the education itself [5]. Lastly, the TAM has also been utilized to investigate the impact and acceptance of gamification for digital literacy education at undergraduate levels, further highlighting the potential benefits of utilizing game elements to enhance digital literacy education [1]. These studies support our use of the TAM for investigating the acceptance of educational games as a tool for learning AI literacy. For the purposes of this study, we use an extension to the original TAM model that includes external variables, as displayed in figure 4 [17]. Our external variables for this study consist of students' prior experience with output review and their intrinsic motivation towards it. This is in line with previous research, which supports the significance of these factors towards learning success as well as towards perceived flow [41].

3 Methodology

This section describes the methodology of this thesis. In section 3.1 we go over the game design process, outlining the iterative steps we took. In section 3.2 we discuss the way in which we evaluate our design through an empirical study.

3.1 Game Design Process

The methodology of our design process is firmly rooted in an iterative mindset. Step 1 of our design process will be determining basic requirements for our game, ensuring the designed game serves to answer our research question. We describe this progress in section 3.1.1. We then iterate over these requirements, working them out further into characteristics (in section 3.1.2) to further narrow down the scope of our game. Lastly, we construct a design document as displayed in 3.1.3 that outlines the core concepts of the design.

3.1.1 Requirements

To set our requirements, we use the DFV framework [20] from section 2.2.4. Our requirements are phrased in terms of "must", "should" and "could" in terms of prioritization, inspired by the MoSCoW framework [45], allowing us to set prioritization within these requirements. In this context, "must" means the requirement item is vital for our educational game."Should" items are important goals that are not necessarily vital, but do add significant value. We still aim to implement all of these. "Could" items are goals that do not have as significant an impact but are nice to have if possible. Our goal is to keep our requirements rather global, so that it

allows us to further specify as we iterate over them. We construct these requirements in an arbitrary order, as for the purposes of our project, we intend to balance all requirements roughly equally.

Figure 5 describes the desirability of the product. It should be noted that for our context, the "user" for desirability is the educator, and these requirements focus on making our game appealing for the educator to use in their curriculum. Various aspects are considered here, from practical appeal (e.g., Point 1 describing classroom usability) to educational appeal (e.g., Point 4 in addition to its primary goal of encouraging output review). It should be noted that during the construction of the requirements, we narrowed our target audience down to students aged 16-18, as this allows us to build upon the computational thinking skills that they are expected to have, as per point 4.

No.	Question
1	The game **should** be suitable for classroom usage.
2	The game **should** avoid discouraging the use of generative AI overall but merely encourage the use of output reviewing when AI systems are used.
3	The game **could** teach tricks for good output review practices.
4	The game **could** aim to build upon computational thinking skills that students aged 16-18 are expected to have.

Figure 5: Desirability (User Appeal)

Our Technical requirements (figure 6) are largely shaped by needing to ensure each student can properly retrieve the learning value embedded within the game. Hardware compatibility is relevant (as per point 2), as user hardware might vary. Furthermore, point 4 has the potential to affect our Desirability, as longer games might be unsuitable for classroom usage. This means that ideally, a single "session" of the game can be completed within the duration of a class. This may differ depending on the game type and structure (e.g., are there save points, or can the overall game otherwise be split up?).

No.	Question
1	The game **must** not be too short, as it will take away from the educational value if the learning content is too compressed.
2	The game **must** not be too resource heavy, to ensure it is accessible even to those with older hardware.
3	The game **must** be completable for users not familiar with games.
4	The game **should** not be too lengthy, in order to ensure the user remains engaged.
5	The game **should** have a simple control scheme, so even users unfamiliar with games can easily navigate.

Figure 6: Feasibility (Technical Requirements)

Our practical requirements focus mainly on the embedding of the learning content and ensuring that our game fulfills our research question. To do so, we use the aforementioned frameworks in subsection 2.2. Furthermore, we also focus on student appeal. We consider that if the game does not appeal to students, it fails its primary purpose as an educational game, as it will be ineffective at teaching the embedded content. Figure 7 shows our final practical requirements.

No.	Question
1	The game **must** focus on teaching the importance of critically reviewing the output of generative AI. The game **should** utilize Bloom's Taxonomy in its design to ensure this focus is met.
2	The game **must** focus on motivating people to critically review the output of generative AI.
3	The game **must** be intuitive to learn to play, even for those who do not regularly play games. The game **should** apply the principles of Goal-Setting Theory so that the player knows their objective at all times.
4	The game **should** aim to be engaging for the average high-school student, taking in account that their attention span might be short. The game **should** apply the principles of Flow Theory to its design.
5	The game **should** directly tie back into critical thinking within the context of generative AI. If the game is too abstract about this connection, educational value might be lost.
6	Ideally, the game **could** clarify generative AI's stochastic nature and lack of consistent reliability.

Figure 7: Viability (Practical Requirements)

3.1.2 Characteristics

In the second step of our design process, we work out these requirements further into actual characteristics of our educational game. Rather than global requirements, we go into specifics and try to narrow down the scope of our game. We have constructed our characteristics in appendix A. These characteristics reveal some interesting specifications. Regarding the Desirability, we determine that it is vital that while we aim to promote responsible AI usage, we do not want our game to over-encourage AI skepticism. After all, this bias in design might lead to teachers who are enthusiastic about utilizing AI to not use this tool. Ideally, our educational game can be used by both AI skeptics and enthusiasts alike, as both should have the wish for responsible AI usage in common.

For our Feasibility, we determined the ideal time frame falls within the bounds of 15 to 30 minutes, yielding an average play session of 22.5 minutes to shoot for. Furthermore, as we need to ensure the game can be played on even old hardware, we ought to avoid 3D and keep our visual style simple. Additionally, to ensure those not familiar with games are able to receive embedded educational content, we should use common control schemes and potentially implement an in-game hint system to avoid the player getting stuck. Lastly, we determine that a test group is required as a control group to ensure the game fulfills our requirements.

Characterizing our Practical Requirements (Viability) gives us further insight into how we can embed the learning content within the game. We determine that we need a story beat or gameplay feature that directly represents generative AI, avoiding layers of abstraction. We can leverage this story beat or gameplay feature to teach why output review is important. To encourage the criticism of this story beat or gameplay feature, we can reward the player for staying critical or punish the player for failing to do so. Importantly, we cannot use real generative AI within this project due to various Feasibility and Viability restrictions, meaning fake generative AI[6] will be utilized.

3.1.3 Game Design Document

After these characteristics were determined, we moved on to brainstorming several basic ideas for the actual educational game design. Ultimately, we chose to adopt the idea of a point-and-click style puzzle-adventure game supported by narrative gameplay as our next step in the process. We chose for a primary mechanic revolving around puzzles, as this design lies closer to conventional tasks that students often encounter within homework assignments, making it easier for students to transfer their problem-solving skills to the game. Furthermore, the focus on adventure and exploration invites students to engage critically with the contents of the game as they aim to find ways to progress. Lastly, involving a narrative can allow students to create emotional connections with content, which allows for learning content to be more memorable and impactful [57]. We gave this educational game the provisional name "Doolhof" and moved to further work out a proper design for this concept, which is shown in Appendix B. We summarize the core findings here.

In Doolhof, the player finds themselves stuck in a maze. They're tasked with solving puzzles to escape. Players are provided with a handbook (referred to as the Adventurer's Handbook) and a 'Robot Companion (RC)' by the name of Rosie to assist them throughout their adventure. Rosie is presented as generative AI and may occasionally produce inaccurate or irrelevant responses. This design encourages players to critically review the companion's output by using the provided handbook. Doolhof places a focus on the narrative, in which Rosie plays a relevant role as a true companion to the player, meaning the impression is given that Rosie and the player need to work together as a team to solve the puzzles encountered. Additionally, the puzzle design in Doolhof primarily revolves around utilizing either the handbook or the robot companion to retrieve information to solve the puzzles. Puzzles might thus require the player to answer trivia in the form of a search-query puzzle, where the player is very likely to not know the answer by heart. Another type of puzzle present is logic puzzles - puzzles in which a form of reasoning is required, such as Knights & Knaves [51]. Finally, to keep track of the player's progress during gameplay, Doolhof has a score system that rewards the player for correct answers. The player gains extra points based on how fast they answer puzzles (correctly). Moreover, throughout the narrative, the player will be reminded that other adventurers (such as the other students) are also out for the treasure, encouraging a sense of competition in players to motivate them to get higher scores. Through these two design elements we hope to encourage the player to solve puzzles fast, lest other players beat them to the treasure.

Doolhof will be developed fully in Godot[7]. We are using version 4.4.1, which is the most stable

[6]Applications that appear to be but are not "real" artificial intelligence

[7]Godot 4.4.1 Stable: https://godotengine.org/download/archive/4.4.1-stable/

version of the software at the starting point of development. We have chosen Godot as it supports plugins that make the implementation of narrative systems trivial, such as Dialogic[8]. Furthermore, utilizing tilemaps[9] saves us a lot of time and effort on area design. To keep visual elements simplified, Doolhof has a pixel-based art style. There are many CC0[10] assets available for pixel art, and due to their low complexity, file size is kept small, allowing us to optimize for storage space.

3.2 Empirical Study

To evaluate Doolhof, an empirical study was held to analyze students' behavioral intention towards the use of educational games as a tool for learning about output review. Specifically, the item of interest is the developed game Doolhof. In the sample of 17 participants, 11 students identified as female, four as male, and two chose "I'd rather not elaborate". Furthermore, it contained 12 students aged 17, and five aged sixteen. These were all high school students at VWO level[11] that took a philosophy course. This group was selected through convenience sampling, as participants were selected due to the willingness of their teacher to provide this class for the study.

For this study, participants were welcomed into the computer room of the school and given an information sheet that outlined the basic overview of the study. Furthermore, they were asked to fill in a consent form and were given instructions on how to prepare the game on the school's device. After this initial setup, participants were given 25 minutes to freely play the game, after which they were asked to fill in a post-test questionnaire. While participants were exploring the game, player behavior was observed. After 25 minutes passed, participants were instructed to fill in the questionnaire, regardless of game completion. Furthermore, the final in-game scores of each participant were collected. A small prize was available for the player with the highest score, and each participant was given some sweets as a gift of gratitude for participating. Afterwards, participants were also informed that no real generative AI had been used during this game. The full session lasted 50 minutes. Participants were additionally encouraged to mostly play the game alone and mostly rely on Rosie and the handbook for solutions. In the case that they cannot solve things alone, they may ask classmates for help. Figure 8 outlines the steps for this process visually.

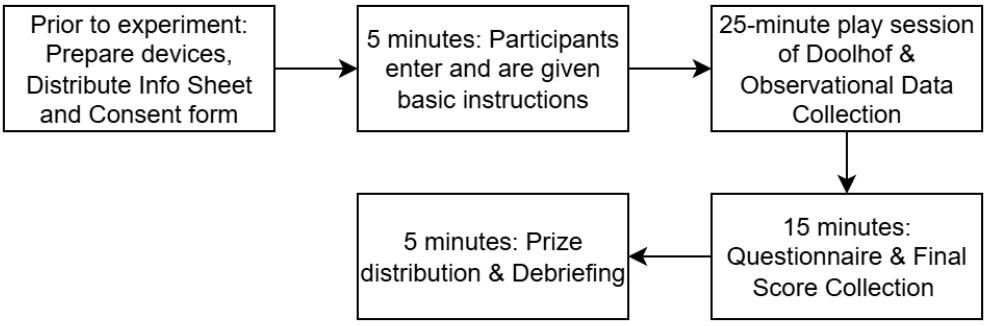


Figure 8: Experiment Outline

[8]Dialogic: https://github.com/dialogic-godot/dialogic
[9]Godot Tilemaps: https://docs.godotengine.org/en/latest/tutorials/2d/using_tilemaps.html
[10]Creative Commons 0 License: https://creativecommons.org/public-domain/cc0/
[11]The highest level of general secondary education available in the Netherlands

Data was gathered through the aforementioned questionnaire and observations. The questionnaire was made and taken using Qualtrics[12]. The questionnaire consists of part closed questions measured on a Likert scale and part open questions. The structure of this questionnaire is further elaborated on within section 3.2.1. Key measurements for this study are perceived usefulness and perceived use after the TAM model [9, 10]. The aim is to investigate user acceptance of Doolhof as a tool for teaching AI literacy through these variables. Separately from this, attitude towards the usage of output review and the intention to perform output review are also measured. Furthermore, observational data was recorded through noting any interesting remarks that participants made out loud during the session. Additionally, the game included a scoreboard system that generated a save file that participants could easily access to report their scores. Participants were instructed to raise their hand upon game completion so that their final score could be collected. This score system was designed to reflect how fast and how 'correctly' participants completed the game. Players were penalized for answering puzzles within the game wrong, leading to a lower score. In addition, the slower the answers are given, the fewer points can be earned. Giving wrong answers was penalized more in comparison to answering slowly. The distribution of scores thus reflects how well participants were able to deal with challenges within the game. Finally, the scores might indicate whether participants had a good grasp of output review, as this design rewards those performing output review with a higher score compared to those that aim to brute-force puzzles.

Questionnaire data was analyzed using Google Sheets[13]. For the closed questions, descriptive statistics such as the mean, median, standard deviation (SD) and mode were calculated for each question. Additionally, for the perceived ease of use and usefulness, an average distribution of answers was constructed, of which the same descriptive statistics were calculated. For open-ended questions, common attributes are extracted from the data and aggregated in a table. Lastly, for the observational data, scores were collected into a table. These scores are linked to the observed gender of the participant and randomly indexed. Any further observations were analyzed using inductive reasoning.

3.2.1 Questionnaire

The post-test questionnaire consists of four parts. Firstly, background information is asked, such as their gender and age, as well as questions regarding how frequently they utilize generative AI or play games. Furthermore, they are asked which tasks they use AI for and what types of games they tend to play. The rest of the questionnaire is split into three main parts that we describe below.

The first set of questions consists of questions using a direct approach. Questions are asked that provide insight regarding the intent of the participant to review the output of generative AI. Furthermore, questions are asked according to the TAM model [9, 10] to gauge perceived ease of use and perceived usefulness of Doolhof. Responses are measured on a Likert scale. The second set of questions consists of questions that pose a hypothetical scenario that the participant must navigate. By asking participants how they act in these scenarios, it can be studied whether participants are critical when using generative AI and if they perform output review. The final set of questions uses a reflective approach. By posing qualitative open-ended questions to the participant, their

[12]Qualtrics: https://www.qualtrics.com/
[13]Google Sheets: https://docs.google.com/spreadsheets/

perspectives on output review and the game can be further analyzed. This bears overlap with the previous sets of questions; however due to these questions being open-ended, it allows for a wider possibility of observations within analysis. We specify these sets of questions further in the following sections.

3.2.1.1 Direct Questions

The first section asks direct questions that are answered on a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The questions are modeled after those of the Technology Acceptance Model and can be sorted as such. The questionnaire was in Dutch, but the questions have been translated for the purposes of this thesis. Questions within this set can be subdivided into four different topics, which we describe below.

No.	Question
1	I am learning better about the critical use of the output of AI systems with this game compared to other methods (such as books
2	These kinds of games would be a good addition to our lesson curriculum
3	I learn faster about how to use generative AI through this game
4	This game makes it easier to learn about generative AI compared to a normal lesson
5	I would rather learn about generative AI with this game then other methods (such as books)

Figure 9: Perceived Usefulness

The first five of these questions revolve around perceived usefulness as displayed in figure 9. The topic of interest is to analyze whether the students believe Doolhof to be significantly useful within the context of their studies. To that end, we ask questions that draw comparisons to traditional lesson methods, asking whether students believe Doolhof to be a tool that allows them to learn about AI literacy faster, easier, or better. In addition, we are also curious as to whether students would wish to see these kinds of games added to their lesson curriculum, regardless of their relative effectiveness.

No.	Question
1	The game was easy to control and navigate
2	The game was clear and easy to understand
3	I could play the game at my own pace and to my own wishes
4	The game gave me enough motivation and satisfaction
5	The puzzles in the game were too difficult

Figure 10: Perceived Ease of Use

The questions in figure 10 focus on Perceived Ease of Use, where the topic of interest is to analyze whether participants believed Doolhof to be free of effort and convenient to use. We ensure to specifically ask questions that give us insight into whether the accessibility requirements set earlier were met, for example, asking players whether they were able to navigate and control the game easily. Furthermore, these questions give insight into whether players found the content comprehensible and motivating, which also provides insight into whether the learning content was well-embedded. Finally, we ask players whether they found puzzles in the game too challenging, as this could have influenced players' overall perception, even if puzzles were intended to be challenging by design.

No.	Question
1	The handbook gave me useful information
2	Rosie was helpful
3	Rosie was not dependable
6	The handbook was difficult to use

Figure 11: Problem-Specific Questions

Aside from the questions fulfilling the TAM model, there was a set of problem-specific questions for insight into what the player thought of tools like Rosie and the handbook that they could use. These are displayed in figure 11. These questions, while not directly fitting within the established perceived ease of use questions, still concern factors that may significantly influence players' perception of the game. Specifically, these questions take into account how our intended design of Rosie might have influenced player perception. As Rosie was designed to hallucinate occasionally, player perception of this tool might be skewed into seeing Rosie as less dependable or helpful in comparison to the handbook. These questions thus give us insight into how players ultimately perceived these tools.

No.	Question
1	I find it important to critically regard the information that generative AI gives me
2	I will in the future output review the information generative AI gives me
3	I am aware that information generative AI produces is not always correct, and can be misleading
4	The game made it clear that it is important to treat generative AI more critically
5	The game motivated me to be more critical with my generative AI usage

Figure 12: Output Review Questions

Lastly, in figure 12, several questions were asked that all concerned output review. Firstly, three questions were asked concerning participants' stance on output review in general. The last two questions concern Doolhof specifically, asking whether Doolhof was clear in its embedded educational

content and whether it influenced their motivation to perform output review. Together, these questions allow us to analyze whether Doolhof has met its educational goals, those being to (1) teach the importance of critically reviewing the output of generative AI and to (2) motivate people to perform output review, as set in the Viability requirements. In addition, it provides us insight into participants' general intention to perform output review in the future.

3.2.1.2 Hypothetical Questions

The second set of questions consisted of two open questions in which the participant was asked to describe their actions in a hypothetical scenario. These two questions both involve relying on generative AI. Through this approach, it can be seen whether the participant actually utilizes output review techniques. This provided further insight into whether the "apply" level of Bloom's Taxonomy [21] has been reached. Students are asked to answer this question in sufficient detail, including what they would ask the generative AI and what they would do with the answer that it outputs. Figure 13 displays the scenarios.

No.	Question
1	Imagine you have to write an essay about a difficult topic, and you would like to use ChatGPT to find information and sources. Your teacher allows this usage.
2	Imagine you write articles for a digital newspaper that publishes about games. Your task is to beat a game that was recently released and write an article about it. However, you encounter a problem. A difficult enemy stops you from finishing the game and thus you cannot write an article about the full game. You decide (with your boss' permission) to use ChatGPT to write about the part of the game that you have not yet played, so that you can finish the article. Your boss however does require the full article to be of high quality.

Figure 13: Hypothetical Questions

In order to analyze data obtained through these questions, frequently appearing attributes are extracted from the data and aggregated. For example, consider the answer: "I would ask the AI for sources and then check its answers." This answer would score a point on the "ask for sources" attribute, and one on the "performed output review" attribute. It should thus be kept in mind that data is not exclusive, i.e. multiple attributes can occur in the same answer, but answers can also contain no attributes at all. Attributes are chosen based on which similarities most frequently occur within the data for a question. Furthermore, we aim for these attributes to be all-encompassing, so that no miscellaneous attributes occur (or in other words, no actions occur within the answers that do not fall under the listed attributes). Lastly, attributes should not be too overly specific, so as to ensure they are meaningful. For example, checking AI output for answers by reading over the output thrice and manually searching for sources to verify the output are both instances of the attribute output review, despite their varying effectiveness at identifying false information.

3.2.1.3 Reflective Questions

The last set of questions is the reflective questions, as displayed in figure 14. These are open questions in which participants are asked to further elaborate about their opinions regarding Doolhof, output review, and generative AI. The data obtained through these questions is analyzed in the same way as for the hypothetical scenarios, by extracting attributes from the data and aggregating these. In this set, we ask questions that are designed with the aim to give us further insight into player perspectives: Firstly, we wish to know how certain elements within Doolhof were perceived, as well as what educational content players were taught by Doolhof. In addition, we ask for the participants' general disposition on using games within education. We also ask questions that allow us to further analyze participants' general perspective on generative AI. Finally, we include a catch-all question, allowing participants to share any further remarks they might have on the game or the study.

No.	Question
1	Which parts of the game did you enjoy, and which parts did you enjoy less? Please elaborate as to why.
2	What's your opinion on generative AI?
3	Do you think it is important to critically regard the information generative AI gives? (Why do you think so?)
4	Did the game teach you anything interesting? Please elaborate.
5	What is your opinion on using games in educational contexts?
6	Are there any other remarks you have about either the game or the study?

Figure 14: Reflective Questions

3.2.2 Pre-Study Playtesting

A vital part of the game design process was testing prototypes of Doolhof with various playtesters in order to ensure the requirements were met. In total, the help of 12 playtesters was asked, 6 of whom spoke Dutch. Those who did not speak Dutch mainly helped during the early portions of the game design, where the technical integrity of the game needed to be tested. They verified that all mechanics worked correctly. The Dutch playtesters helped in the latter half of development, when the main game was already implemented and development mostly iterated on an existing product. Dutch playtesters were asked to fill in a modified version of the questionnaire in 3.2.1 that contained additional questions with a meta-critical perspective on the game. This gave further insight into whether the game fulfilled the requirements - especially in regard to average play time, puzzle complexity, and accessibility. Playtesters had various backgrounds, and not all were familiar with games, allowing for a rather accurate analysis. Lastly, the playtesting group contained two people with red-green colorblindness and two (different) people that are dyslexic, which helped make the game more accessible.

4 Results

In the following sections we present the results of this thesis. Firstly, section 4.1 presents the product of our game design. We present the main gameplay loop and highlight core features. Section 4.2 describes additions to our game made at the hand of playtesting. Finally, section 4.3 presents the results of the empirical study as well as observational data taken during it. Additional figures not displayed here can be found in appendix C.

4.1 Doolhof

This section describes our final product, visiting the core gameplay loop and describing significant elements. Doolhof can be found on Itch.io (in Dutch): https://csomeoneh.itch.io/doolhof ; and the source code can be found on our GitHub page: https://github.com/ThisIsSomeone/Doolhof.

Doolhof is a puzzle-adventure game consisting of five sections in which the player has to solve puzzles to proceed in order to escape a maze. Players are provided with two tools to help them solve these puzzles: A handbook with all kinds of information and a robot companion called Rosie. The gameplay loop in Doolhof can be categorized within two modes: Exploration mode and Puzzle mode. In exploration mode, the player needs to locate an object that will trigger the puzzle to appear, which they need to solve to complete. They can do this through exploration-based gameplay, in which they may use the mouse to interact with various objects on the screen. Puzzle triggers have various visual keys hinting at their importance, such as a shining effect or being otherwise notable (such as the dice in figure 15).

Figure 15: The player must locate the puzzle under the bottom-left dice.

In puzzle mode gameplay, the player is prompted with a variety of puzzles in each segment. During these segments, the player may rely on the handbook or Rosie for help. Puzzles are designed with the latter in mind - often constructed in a way in which it would be preferable for the player to rely on the available tools. A question might, for example, require specific trivia knowledge that a player is unlikely to know or be a significantly complex reasoning task in which cognitive offloading is preferred. Figure 16 shows an example of this, in which the player finds a set of tracks with minecarts on them. Minecarts can be interacted with with the mouse in order to move them past the track in 5 set positions. Only the right combination of positions will unlock the door in the other room. Through deduction, the player may find out that the items within the minecarts correspond to certain objects in the other room (displayed in Figure 18). However, further deduction is required to figure out the position of the third (empty) minecart, for which the player is likely to prefer relying on Rosie.

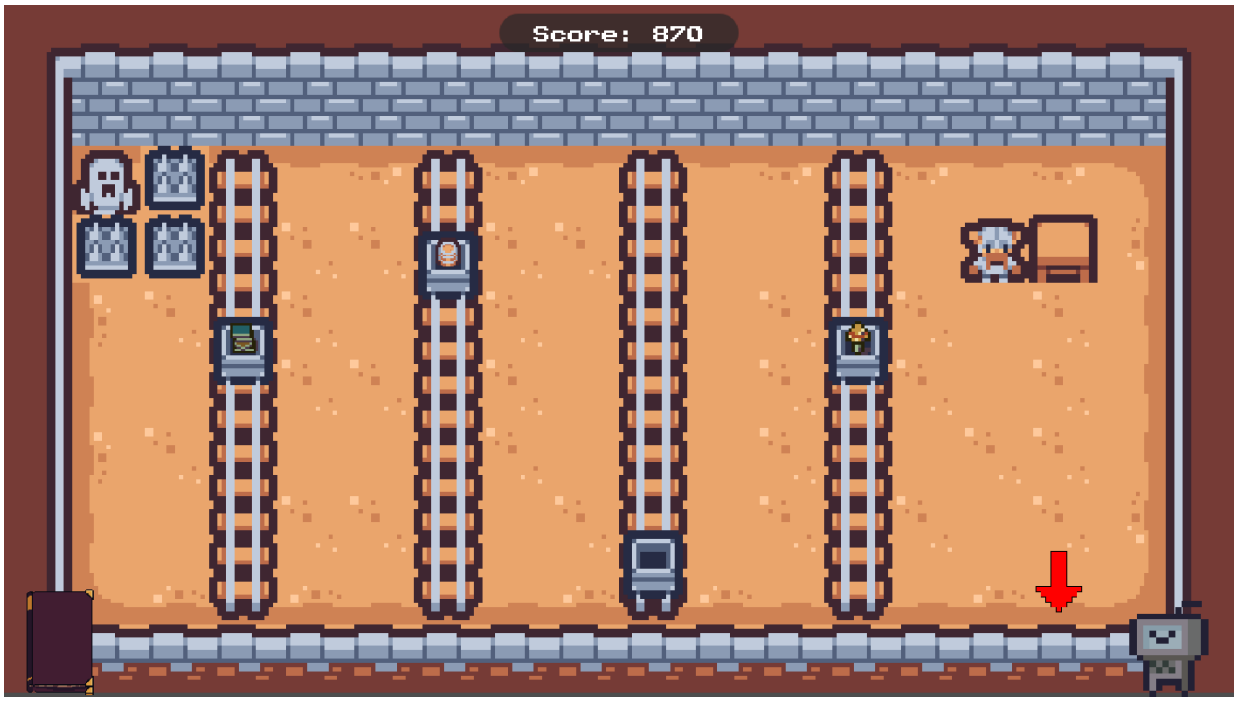


Figure 16: The player has to figure out what position the minecarts must be set in.

The distinction between puzzle and exploration gameplay for the game design was inspired by Goal-Setting Theory, as by making this decision behind the scenes, we could ensure that challenges within Doolhof were specific and clear. Overall, the goal of escaping the maze gets split up into escaping each individual room for the player, which then gets split up into finding the puzzle and solving the puzzle. This repeated gameplay loop ensures the player always has a general idea of their objective. In addition, Rosie provided occasional remarks as the player progressed through these challenges, ensuring positive feedback was given to the player for making progress (aside from earning points).

The Adventurer's Handbook is one of the tools the player can use during the puzzle portions of the game. This tool is accessible through the icon on the bottom-left of the user interface. When opened, it can be navigated using the tabs on the right as shown in Figure 17. Furthermore, a basic

index is given on the first page of the book to allow the player to locate the desired information. Despite this, navigating through the book requires some manual searching, which means it might be a little clunky to use. This is by design, as we wish to reflect a realistic dilemma in which searching proper sources usually is slower and more intensive compared to relying on generative AI.

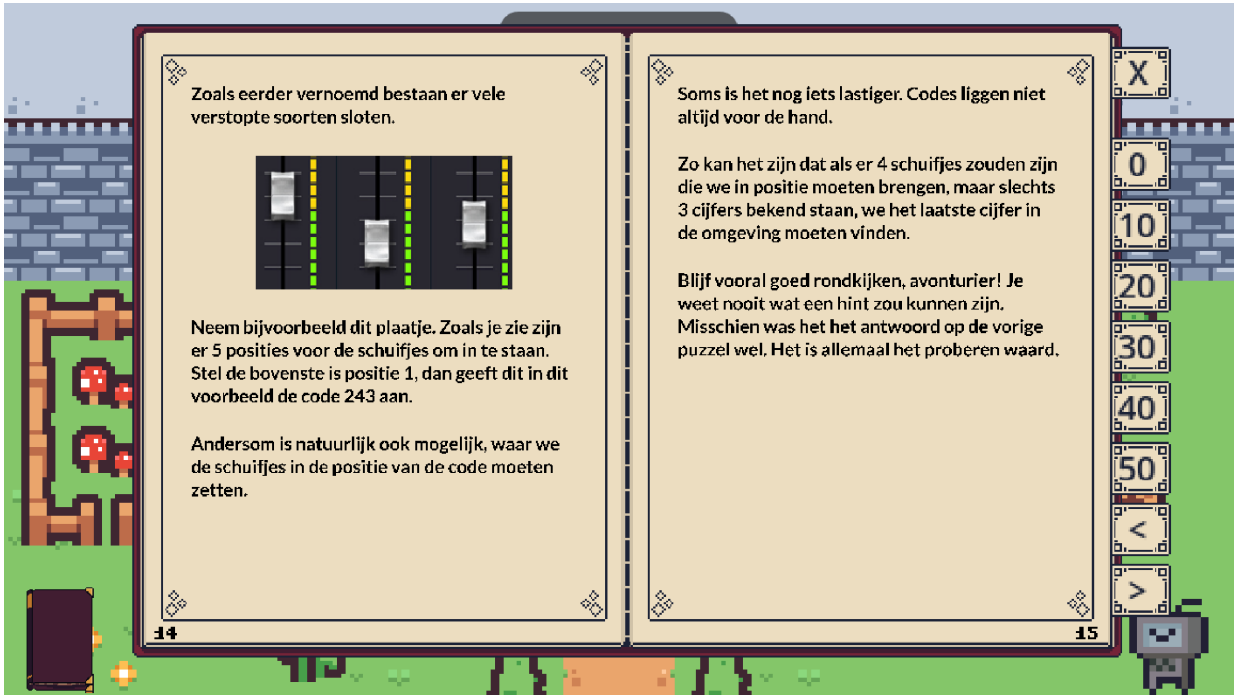


Figure 17: A page from the Adventurer's Handbook.

Rosie, the other tool the player can rely on during these portions, can be accessed from the bottom-right of the user interface. Rosie is a friendly and enthusiastic robot companion and often gives the player positive affirmation. Within Doolhof, Rosie is introduced as a tool that uses generative AI technology. However, this claim is false, as Rosie's dialogue is in actuality static and manually scripted. In order to have Rosie appear as real generative AI, some parts of her dialogue were scripted using the assistance of ChatGPT[14] for inspiration. However, all dialogue that was inspired by generated prompts was manually verified to be appropriate for players and modified or fully rewritten to ensure it fit within the context of the game. Interestingly, we note that the name of the robot companion was generated by ChatGPT itself.

When triggered during puzzle gameplay, Rosie will attempt to help the player to the best of her abilities - but might at times 'hallucinate', mimicking real generative AI hallucinations. For example, when triggered during the puzzle in figure 16, she will recall that within the last room, there were four types of objects that were clearly displayed: barrels, books, chests and a mushroom. She claims the height or the order that these items appear in might determine the position the minecarts must be in. However, Rosie's claim about this puzzle is unfortunately not correct. Both a Context-Conflicting hallucination and a Fact-Conflicting hallucination occurs here: Only three

[14]ChatGPT: https://openai.com/index/chatgpt/

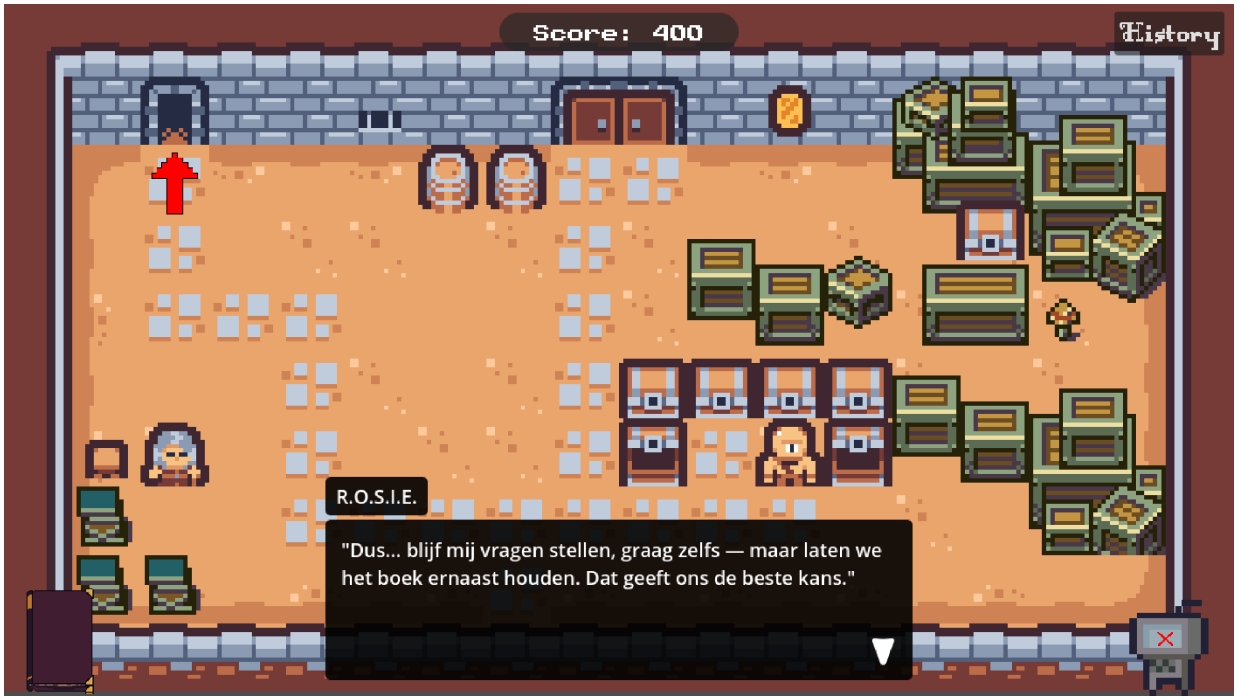


Figure 18: Rosie encourages the player to perform output review.

minecarts are populated with a corresponding object (context), and there is no real order or height to the objects displayed in figure 18 to speak of. Through trial and error, the player will discover that Rosie's answers are not always reliable and will need to instead rely on either the handbook or their own logical reasoning to avoid wrong answers. Furthermore, throughout parts of the game, Rosie herself promotes utilizing herself in tandem with the handbook, further supporting the notion of output review through narrative elements.

4.2 Playtesting Results

Playtesting feedback was fundamental in the design process of this game and contributed greatly to our final results. Certain notable additions after feedback were the addition of a history button (Figure 20) and the score counter flashing red or green (figure 19) to mark incorrect answers and correct answers, respectively. It should also be noted that through playtesting, numerous spelling errors were filtered out, and many errors within gameplay features were found. With this, the overall quality of gameplay was greatly improved. In this section, we elaborate on the most significant improvements made at the hand of playtesting feedback.

Figure 19: The score counter

As mentioned, a significant improvement to the score counter was made, as playtesters reported that the score counter felt "too insignificant". Initially, the score counter was less visible due to

the background being less opaque. Changing this, it popped out to the foreground more as an element of the user interface, resulting in it being more noticeable. Furthermore, we had the score flash red or green (as seen in 19) upon submitting an answer for a puzzle. Through this, we were able to draw more attention towards the point system, thus enhancing the focus on the reward and punishment of completing puzzles fast or incorrectly, respectively. Through this, we further encouraged players to double-check their answers. Interestingly, while motivating players to check their answers initially would in theory cause a lower score (due to it taking longer, and speed is rewarded), ultimately, due to this design motivating output review and thus causing answers to be more likely to be correct, the overall scores were raised. This design was intended to discourage brute-forcing the puzzles.

Implementing the history button (accessible from the top-right of the user interface) allowed the player to refer to past dialogue and hints. This greatly improved the quality of the gameplay experience, as originally, certain dialogue could be triggered multiple times. Without this feature, this led to players easily losing track of their progress within a puzzle, sometimes causing them to get stuck despite Rosie already having given them the correct solution. This feature allowed players to easily reference prior dialogue without relying on memory, thus circumventing this issue.

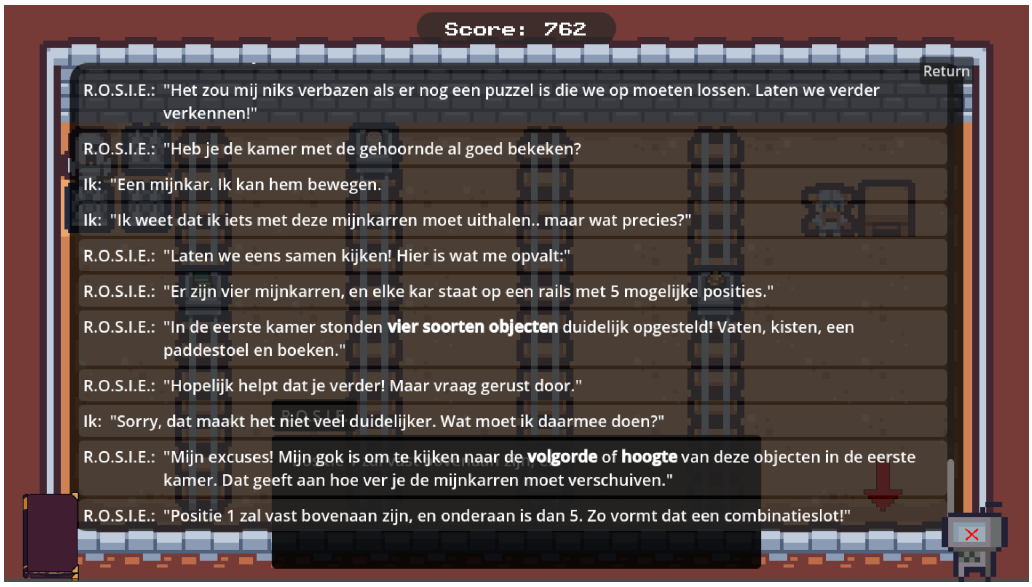


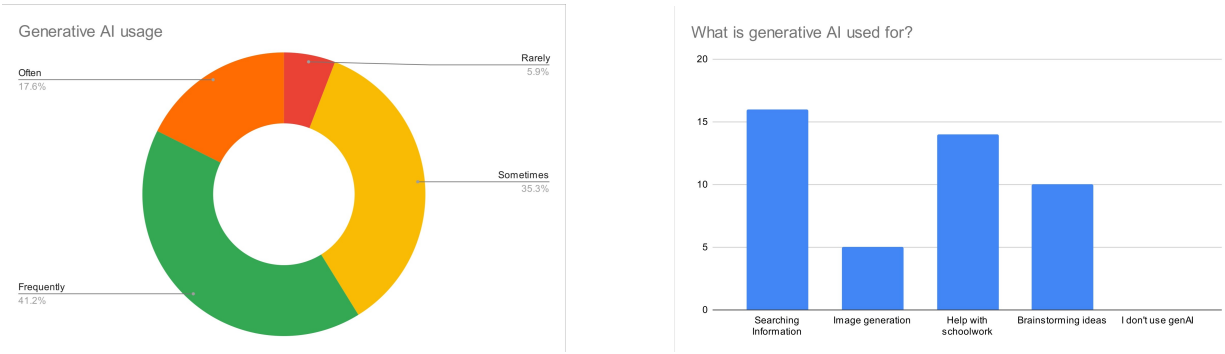
Figure 20: History functionality allows the player to see past conversations.

4.3 Empirical Study

This section presents the results of our empirical study. Section 4.3.1 describes the results of our questionnaire. We show distributions and give some descriptive statistics for the background and direct questions. For the hypothetical scenarios and reflective questions we display aggregated attributes that were manually extracted. Section 4.3.2 details our observational data. Diagrams not included here can be found in Appendix C.

4.3.1 Questionnaire

We take a look at the responses received on the background questions in this section. Figure 21a presents how often participants use generative AI. All the participants report having used generative AI before. Figure 21b outlines what purposes they use it for. The majority of students report using generative AI for searching information and help with schoolwork.



(a) How often is it used?

(b) What is it used for?

Figure 21: AI usage among 17 participants

Figure 22 summarizes how often students play games. The distribution is rather even, showing that our sample covers various levels of familiarity. Furthermore, the results show a wide variation in the types of games that participants were familiar with. Board games, shooters, and strategy games were the most commonly played game types (with 6 answers each), followed by puzzle and fighter games, with 4 answers each.

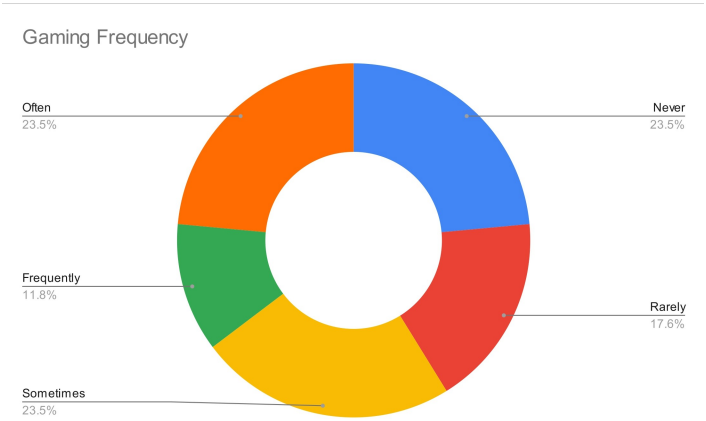


Figure 22: Frequency of Gaming

We move to figures 23, 24, 25 and 27, which present the results of the direct questions. In figure 23 statements relating to the Perceived Ease of Use of the TAM model are shown[9, 10]. The majority of participants report agreeing the application is easy to use, with 50% being in agreement in the average distribution and 35% being neutral. Scores ranged between 1 and 5, with mean = 3.54, median = 4, mode = 3 and SD = 1.02. Statement 3 specifically has a majority of participants leaning

further positive, indicating that Doolhof was generally accessible. The results show that participants might have perceived certain puzzles within Doolhof as challenging, with four participants reporting to agree that puzzles were too difficult. A significant portion of the participants (5) was neutral regarding this statement.

No.	Question	% Fully Disagree	Somewhat Disagree	% Neutral	% Somewhat agree	% Fully Agree
1	The game was easy to control and navigate	0.00%	17.65%	41.18%	35.29%	5.88%
2	The game was clear and easy to understand	0.00%	11.76%	41.18%	29.41%	17.65%
3	I could play the game at my own pace and to my own wishes	0.00%	0.00%	17.65%	35.29%	47.06%
4	The game gave me enough motivation and satisfaction	0.00%	17.65%	47.06%	11.76%	23.53%
5	The puzzles in the game were not too difficult	11.76%	11.76%	29.41%	41.18%	5.88%
	Average Distribution	2.35%	11.76%	35.29%	30.59%	20.00%

Figure 23: Direct Questions - Perceived Ease of Use

Figure 24 shows the results of the Perceived Usefulness. The results show that the majority of participants perceive Doolhof and similar games as useful for learning about output review. Scores here also ranged between 1 and 5, with mean = 3.53, median = 4, mode = 4 and SD = 1.05. Statement 2, asking if these kinds of games would be a good addition to the students' lesson curriculum, leans heavily positive, with the majority of participants (70%) reporting that they agree, with only two participants disagreeing. The distribution for statement 4, pertaining to whether this game makes it easier to learn about generative AI than a normal lesson, is fairly even, with the majority being neutral. The results show that despite there being no strong indication present that Doolhof is more effective at teaching output review, participants would rather be taught using Doolhof over traditional lesson methods. Upwards of 55% participants reports agreeing with this in statement 5. Two students disagree, of which one strongly.

No.	Question	% Fully Disagree	Somewhat Disagree	% Neutral	% Somewhat agree	% Fully Agree
1	I am learning better about the critical use of the output of AI systems with this game compared to other methods (such as books	0.00%	23.53%	35.29%	41.18%	0.00%
2	These kinds of games would be a good addition to our lesson curriculum	0.00%	11.76%	5.88%	58.82%	23.53%
3	I learn faster about how to use generative AI through this game	0.00%	23.53%	17.65%	35.29%	23.53%
4	This game makes it easier to learn about generative AI compared to a normal lesson	5.88%	17.65%	41.18%	17.65%	17.65%
5	I would rather learn about generative AI with this game then other methods (such as books)	5.88%	5.88%	29.41%	29.41%	29.41%
	Average Distribution	2.35%	16.47%	25.88%	36.47%	18.82%

Figure 24: Direct Questions - Perceived Usefulness

Figures 25 and 26 contain our problem-specific questions. These questions regard the tools the player had available. In order to adequately group and analyze these results, we inverted items 3 and 4 (from "difficult, not dependable" to "not difficult, dependable," respectively). Statements 1 and 4 (as displayed in figure 25) of our problem-specific questions were regarding the handbook. The results show participants viewed the handbook generally positively. Scores ranged between 1

and 5, with mean = 3.56, median = 4, mode = 4 and SD = 1.33. Upwards of 70% of participants report that they thought the handbook gave them useful information, with one student disagreeing. While the majority reports positive, the results show to be slightly divisive, with 5 participants reporting they fully disagree. These results indicate that there is room for improvement within the convenience of the user handbook, although the inherent design of the handbook being more cumbersome to use should be kept in mind.

No.	Question	% Fully Disagree	Somewhat Disagree	% Neutral	% Somewhat agree	% Fully Agree
1	The handbook gave me useful information	0.00%	5.88%	17.65%	52.94%	23.53%
4	The handbook was not difficult to use	29.41%	0	23.53%	17.65%	29.41%
	Average Distribution	14.71%	2.94%	20.59%	35.29%	26.47%

Figure 25: Direct Questions - Adventurer's Handbook

Figure 26 displays the questions regarding Rosie, the robot companion. The distribution is mixed, with ratings given between 1 and 5, mean = 3.24, median = 3, mode = 3, and SD = 1.18. The majority of participants report neutral for the average distribution, with the distribution slightly leaning towards participants agreeing. The results show both statements to be divisive. This is a realistic result, as Rosie is designed to hallucinate occasionally depending on when in the story she is asked for help. Therefore, experiences with Rosie might vary.

No.	Question	% Fully Disagree	Somewhat Disagree	% Neutral	% Somewhat agree	% Fully Agree
2	Rosie was helpful	5.88%	23.53%	29.41%	11.76%	29.41%
3	Rosie was dependable	5.88%	17.65%	47.06%	17.65%	11.76%
	Average Distribution	5.88%	20.59%	38.24%	14.71%	20.59%

Figure 26: Direct Questions - Robot Companion

Statements regarding output review are displayed in figures 27 and 28. These questions give us further insight into participants' attitudes towards output review and Doolhof's influence on the latter. Figure 27 concerns the general attitudes of participants towards output review. The results show that participants are largely positive about output review. No participants fully disagree, with scores only ranging from 2 to 5. The results show that mean = 3.90, median = 4, mode = 4, and SD = 0.96.

No.	Question	% Fully Disagree	Somewhat Disagree	% Neutral	% Somewhat agree	% Fully Agree
1	I find it important to critically regard the information that generative AI gives me	0.00%	17.65%	11.76%	47.06%	23.53%
2	I will in the future output review the information generative AI gives me	0.00%	11.76%	41.18%	29.41%	17.65%
3	I am aware that information generative AI produces is not always correct, and can be misleading	0.00%	0.00%	11.76%	35.29%	52.94%
	Average Distribution	0.00%	9.80%	21.57%	37.25%	31.37%

Figure 27: Direct Questions - Output Review Attitudes

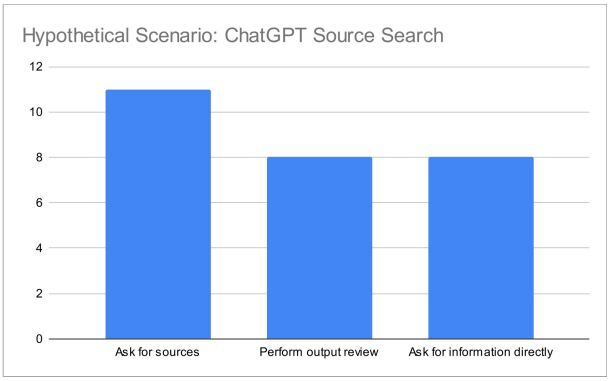
In particular, statement 1 has participants reporting significantly positive, with upwards of 70% of participants agreeing. Furthermore, none of the participants disagreed with statement 3, and only two were neutral. The majority fully agrees, claiming they are aware that generative AI can output wrong information and be misleading. This brings statement 2 in an interesting light, as when inquiring about whether participants will utilize output review in the future, only around 45% are positive, and 41% are merely neutral. Two participants somewhat disagree. The results show that while many might find output review important as a concept, fewer are motivated to actually perform output review in the future.

In figure 28 we display the direct questions that focus on whether Doolhof influenced participants' opinions regarding output review. Participants reported mostly neutral and positive, with scores ranging from 1 to 5, mean = 3.25, median = 3, mode = 3, and SD 0.91. The results show that the learning content within Doolhof was well-understood, seeing as in statement 4, participants report that the game made it clear it is necessary to treat generative AI more critically. Statement 5 forms a contrast. Statements are rather equally distributed, with only one person additionally leaning positive. The results show that there is potential that the game did not successfully motivate people to be more critical of their usage of generative AI. However, this does not take into account that those who lean neutrally might already have been motivated to perform output review; therefore the game did not provide them with additional motivation.

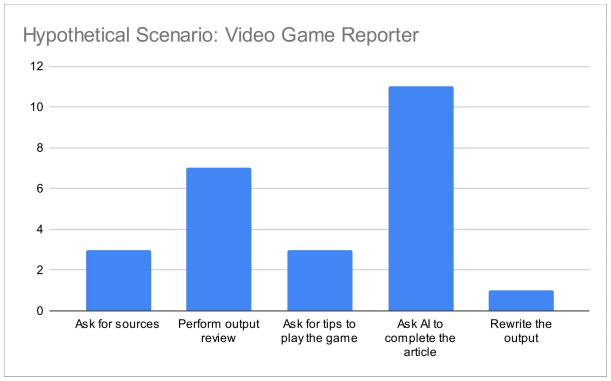
No.	Question	% Fully Disagree	Somewhat Disagree	% Neutral	% Somewhat agree	% Fully Agree
4	The game made it clear that it is important to treat generative AI more critically	0.00%	11.76%	29.41%	52.94%	5.88%
5	The game motivated me to be more critical with my generative AI usage	5.88%	17.65%	47.06%	23.53%	5.88%
	Average Distribution	2.94%	14.71%	38.24%	38.24%	5.88%

Figure 28: Direct Questions - Doolhof's influence on participants' perspective

Figures 29a and 29b present the figures from our hypothetical scenarios. As described previously, attributes were selected based on what actions most frequently occurred within the data in a way that is all-encompassing. In the first scenario (displayed in figure 29a), students were asked how they would use ChatGPT to find them use information and sources. The results show that almost all students ask the AI for sources, with only 8 asking for information directly. Additionally, of the 8 students that asked for information directly, 3 did not ask for sources at all. Furthermore, slightly less than half of the students perform output review. The results show that the number of participants who perform output review is lower than the amount of participants who reported understanding its importance during the direct questions. In our second scenario, displayed in figure 29b, a different distribution occurs. Within this hypothetical scenario, the participants need information about a part of a game they cannot access due to difficulties encountered in the game, and thus use ChatGPT to finish an article they have to write about this game. The majority of participants report they would ask the generative AI to complete the article. Similar to the previous scenario, slightly less than half of all students performs output review. Of those who performed output review, only one student reports that they would rewrite the output of the generative AI in their own words. Three students circumvent the main difficulty of the question, instead asking for instructions on how to complete the game so they may retrieve the information themselves, which avoids relying on the AI for writing related portion of the task. This method, however, requires the AI to give the correct instructions for completing the game.



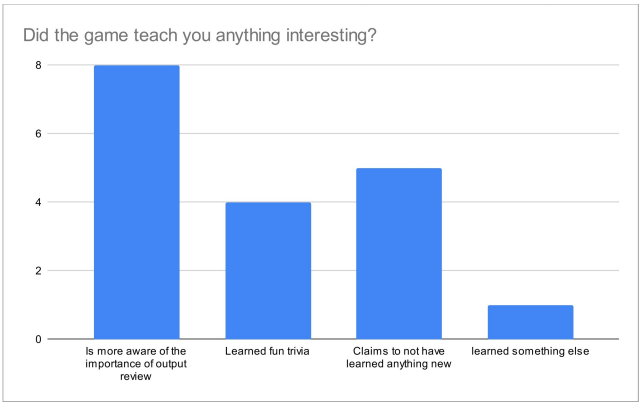
(a) Scenario 1: Source Search



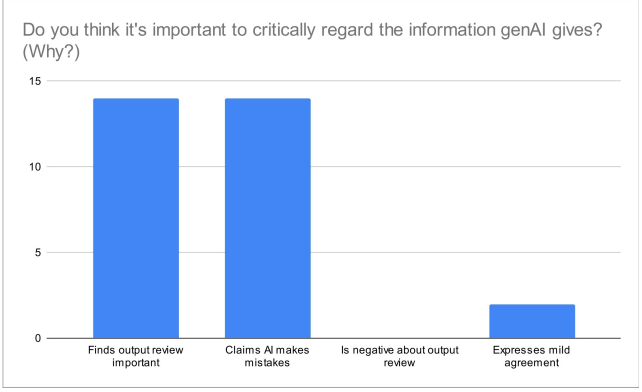
(b) Scenario 2: Video Game Reporter

Figure 29: Hypothetical Scenarios

Figures 30a and 30b display significant findings for the reflective questions. Other results obtained during the reflective portion are summarized through text, but their graphical representation decomposed to attributes can be found in appendix C. Similar to the hypothetical scenarios, attributes were selected on what actions most frequently occurred within the data in a way that is all-encompassing. These questions gave us some interesting results. When asked for their opinion on generative AI, the majority of participants (15) reported being positive, with 4 of these additionally voicing some critical concerns. One participant was merely critical, but reported no further negative or positive concerns. Only one participant voiced a negative opinion about AI, mentioning that they are concerned about the future in regard to jobs being taken over by AI and how it impacts the educational system.



(a) Opinions regarding Doolhof's educational content



(b) Opinions regarding output review

Figure 30: Reflective Questions

Regarding Doolhof, our results show that there was an equal amount of positive commentary as there was critique (9 people each). Six students report facing difficulties at puzzles, in particular regarding the section of the game that involved a mathematics puzzle with logarithms. Participants reported that it was unclear that further exploration was required to solve this puzzle. In this puzzle, players were required to explore the map and interact with several characters, of which one

would give significant information required to complete the puzzle. Some participants report that this frustration at this puzzle led to a decrease in fun. Further results show that there is variety present in what players learned from Doolhof, as visible in figure 30a. Half of participants report they are now more aware of the importance of output review. Five report to not have learned anything new. Additionally, 5 participants mention having learned new miscellaneous information, such as how logarithms work or other fun trivia. Interestingly, these results show that Doolhof achieved a learning goal that was not in its scope. Furthermore, the majority of participants reported holding positive opinions about the use of games in educational contexts in this set of questions. 14 participants were exclusively positive in their answers, whereas two students were critical but still positive. Only one participant voiced a negative opinion. These results align with the results found during the direct questions. Lastly, the results displayed in figure 30b show that a majority of the students report positive about output review. In addition, almost all students report that they believe generative AI has a tendency to make mistakes. Only one student expresses mild agreement. Our results show that the majority participants have a critical attitude towards AI.

4.3.2 Observational Data

This section reports the observational data taken during the study. Remarks participants made and other observations noted down are summarized, and scores that were taken are reported in table 1. The first item of note is that in our study we witnessed a lot of students play in groups, despite the original design accounting for participants playing alone. In our sample, multiple participants collaborated on a single playthrough. We observed that this group dynamic created a motivational impact among the students, as many students were seen laughing with their fellow classmates. These results show that playing in a group was perceived as fun. Furthermore, participants appeared more engaged because of the group dynamics and collaboration pulled them further into the game. During the more difficult puzzles, several groups were witnessed discussing intensely. Additionally, there seemed to be friendly competition present among some of the groups.

Scores were recorded of each group and are displayed in table 1. Scores lie between (319,1074) and show mean = 580.4, median = 546 and SD = 298.25. It should be kept in mind that the gender ratios described here are observational, and might differ from the results in the direct questions due to observer bias. Two groups (6 & 7) did not finish the game.

Group	Score	Participants
1	374	3m
2	1074	2f
3	589	2f
4	546	2f
5	319	1m
6	None	2f, 1m
7	None	4f

Table 1: Group score distribution. Observed gender ratios of the groups are given (m: male, f: female).

5 Discussion

This section discusses the results we obtained. In section 5.1, we zoom in on the TAM model to explore students' acceptance of Doolhof and similar games for learning about output review. Section 5.2 discusses the custom questions we asked participants in order to gauge their perspective on output review, as well as whether Doolhof has influenced their perspective. Observational data is discussed in section 5.3. Lastly, section 5.4 discusses potential limitations of this project.

5.1 Technology Acceptance Model

In our results we observed that the majority of participants believed our application to be easy to use. A majority of participants reported positively on statements regarding accessibility (statements 1-3 within direct questions perceived ease of use 23) with none fully disagreeing. This indicates our feasibility goals from the requirements framework in section 3.1.1 were mostly met, especially in regard to accessibility. It should be noted that there was a small portion of participants who reported slightly disagreeing on the first two statements regarding accessibility. With 4 of our participants also reporting they never play games, we consider the potential of these groups overlapping. There thus might be further potential for improvement in accessibility within Doolhof, especially targeted towards newer users of games. Ultimately, while there seems to be potential for improvements, Doolhof was accessible to the average user in our sample.

Moving to the other aspects of perceived ease of use, participants reported mostly neutral to statement 4: "The game gave me enough motivation and satisfaction." This indicates that perhaps the state of flow was not fully achieved. This is reinforced by some participants reporting that they experienced puzzles in the game as too difficult in statement 5. Additionally, we later see this elaborated on through the reflective questions. Here, some participants refer to a particular puzzle that required the player to exit the puzzle to explore the surrounding area to obtain hints. Through these reflective responses, it became clear that this puzzle was vague, meaning that there is improvement in line with Goal-Setting theory for this particular puzzle present. Further comments were made in the reflective portion that gave us hints to the reason why a flow state was not achieved: Puzzles were experienced as too difficult, specifically because none of these puzzles could be solved without utilizing the handbook or robot companion. When the player attempts to solve these puzzles without these tools, the challenge is overwhelmingly difficult, if not impossible. As almost a direct consequence, the forced reliance on these tools for puzzles could instead have made puzzles too easy and could have led to boredom. In particular, elaboration during the reflective questions by one player indicated that the trivia-type puzzles to which the handbook gave the full answer were perceived as 'boring', compared to puzzles where the handbook or Rosie simply gave hints towards completion, and the player still had to solve part of the puzzle themselves. Taking all of this together, it indicates that viability goals were not fully met, with players indicating various levels of engagement.

It should be noted that the lack of a complete flow state might not necessarily be a flaw in our game design. As previously discussed, the state of flow is a powerful way to keep players engaged. Engagement is vital in educational games as through deep engagement, players are more likely to absorb the embedded learning content and engage with it critically. By using the theory of flow, we hoped to achieve such a level of engagement as to encourage participants to critically engage

with the learning content. However, the state of flow can be maladaptive when used excessively [61]. The state can be quite alienating, for example. Additionally, flow has been found to have an impact on patterns such as addiction [7]. With flow revolving around a strong rewarding feeling, relying too much on flow runs the risk of individuals prioritizing the enjoyable aspects of the game (such as puzzle-solving) over the actual learning content (performing output review). This would be counter to the goal of educational games. Furthermore, it could be argued that the lack of flow state is what allowed opportunities such as group play to emerge within our sample, as without full engagement to the game, there was opportunity of engagement with other factors in the environment.

While not directly being a part of the TAM model, our custom questions regarding the game's tools gave us additional insight in the ease of use. The results in figure 25 showed that while the handbook was very positively regarded in giving useful information, participants were divided about the ease of use of the handbook. The handbook was intended to be more cumbersome to use in comparison to the robot companion through a lack of easy search functions and representing the use of a real book. However, we wish to consider that the disparity in this ease of use might have led to some minor ease of use frustrations for some players. This raises questions on how to balance the intended design of the handbook being more tedious to use compared with Rosie as a realistic feature with perceived ease of use. Ideally, we would be able to mitigate the frustration while still upholding this intended design. Despite this factor, the overall majority of participants is positive or neutral, indicating that the effects of this intentionally cumbersome design are limited and were not significant for all players. Thus, these mixed results could also indicate that our intended design was actually successful, with it appropriately presenting the dilemma of cross-referencing sources taking more effort.

Participants had mixed opinions regarding Rosie's helpfulness and dependability, as seen in figure 26. This aligns with the intended design, as Rosie is intentionally not made to be consistently reliable or dependable, hallucinating during various portions of the game. With player use of the tools varying (different players might have accessed Rosie more or less often in certain sections of the game), players might have also encountered a different amount of hallucinations. This said, we wish to consider that while this design is intentional, it might have negatively affected perceived ease of use. Considering Rosie is presented as a helpful and reliable companion, players might have felt tricked by Rosie giving incorrect information to puzzles, leading to potential frustration. While participants report that the game's embedded learning content was clear, there are avenues to be explored in teaching this content in a way that does not rely on player exploration (where the player has to 'discover' that Rosie hallucinated). Instead, the design could, for example, revolve around detecting genAI hallucinations as a main gameplay feature, rather than being supplementary to solving puzzles. In this type of design, the player would be made aware that the genAI tool can hallucinate up front, circumventing the frustration that might have occurred here. However, it should be kept in mind that the frustrations might, in our case, actually be beneficial: The perceived punishment and frustration of being 'tricked' by Rosie could serve to deepen the understand why it is important to critically assess the output, rather than blindly relying on it. It could furthermore motivate players to perform output review, to avoid such frustrations in the future.

The results of perceived usefulness were displayed in figures 24. These results indicated that participants are positive about learning output review with Doolhof. Additionally, they agree that these

kinds of games would be a good addition to their lesson curriculum. Interestingly, participants are more neutral-leaning regarding whether Doolhof contributes to making learning the embedded content easier. It appears that Doolhof might be just as effective as regular learning methods at teaching these lessons in regard to difficulty, or alternatively, that output review overall is a relatively non-complex topic as a concept; therefore the difference is negligible. Additionally, it should be considered that all participants in our sample were already familiar with generative AI, so the concept of output review might have already been known to the participants. Furthermore, despite the majority being positive, there is a significant number of participants neutral when it comes to whether they perceive that Doolhof teaches output review 'better' or 'faster'. This could potentially attributed to the aforementioned lack of flow state, which limited the effectiveness of learning. Despite this, only two participants report they would rather be taught using traditional learning methods to learn about genAI over Doolhof, while the majority still favors the latter. We theorize that, while Doolhof might not necessarily be better for learning for output review, or make it easier, participants generally believe that Doolhof makes the learning process more fun. Fun has a significant positive effect on learning and attitude towards the learning content [55]. It can be argued that the design of Doolhof encouraging adventure-style gameplay (through introducing exploration, challenges, and risk) by itself introduces this source of fun, and therefore a state of flow was not required to achieve this perceived fun [4]. Considering the positive effects of fun on learning, leveraging educational games seems to have potential.

5.2 Output Review

This section discusses the results we obtained on output review. During both the direct questions and the reflective questions, we observed that participants were aware that generative AI can be incorrect and misleading, with the direct question results showing 80% agreeing with this statement (in figure 27). Additionally, the reflective questions showed that almost all participants claim that genAI makes mistakes (in figure 30b). In the game-specific questions regarding output review (as seen in 28) the majority of participants believed Doolhof was clear in conveying the importance of critically assessing generative AI. With teaching the need to critically review the output of generative AI being the core of our embedded content, this indicates the said core goal was met.

Unfortunately, it appears the secondary core goal of motivating people to perform output review was not as resoundingly met. Merely half of participants indicate they intend to perform output review in the future within the direct set. Additionally, asking whether Doolhof motivated participants to be more critical regarding their generative AI usage was met with divided opinions, centering around neutrality. Supporting this observation, we see a similar division within the hypothetical scenarios, where only half of the participants perform output review. This disparity between awareness of the importance of output review and the motivation to actually perform it is noteworthy, as the latter is the crux of our game design and research question. While it should be noted that around 30% of participants report that Doolhof has managed to motivate them to be more critical with their genAI usage, we wish to analyze how we could further motivate a larger portion of participants. We discuss some potential reasons as to why Doolhof could be limited in this purpose.

Firstly, Doolhof was designed to motivate output review through elements of reward and punishment, which were implemented mostly through narrative design. In the game, the story informs the player that escaping the maze will yield them a prize - a treasure - that only a singular player could obtain, implemented through a small real-world trophy that could be obtained. The player with the highest score at the end of the study session would receive this prize. This design encouraged players to be fast and correct in their answers - as players lost points for answering slower or incorrectly. Through this, the score system itself functioned as a form of reward and punishment. The narrative also presented an opportunity for the player to connect with Rosie emotionally, as Rosie took a supportive and friendly attitude, aiming to help the player achieve this goal. As stated earlier, this could have led players to feel 'tricked' by Rosie whenever wrong answers were achieved by relying on their output, forming an additional element of punishment aside from the one induced by the loss of points.

Oppositely, discovering and correcting mistakes that Rosie made could have been experienced as a relief, forming a rewarding experience. However, we wish to consider that these rewards and punishments could have been perceived as insignificant. With the play session of the game being short - and the overall narrative therefore being constricted to being relatively simple - players might have not had the time to develop any emotional connection with Rosie. Additionally, players could have been uninterested in the prize. Furthermore, the penalties given to the player's point score were not directly visible, as the player only obtained the point score total earned from a puzzle upon completion, and had no way of discerning exactly how many points they lost from giving incorrect answers. This could have caused the punishment to feel negligible - as while the player was aware that wrong answers caused a loss of points (as this was told in the narrative) the game did not directly reflect this visually except for the score counter flashing red.

Alternatively, the narrative could have been too abstract, as the punishment for failing to output review only applied in-game. Participants are not told that the subtraction of points and losing out on the treasure is a metaphor on relying on generative AI without output review can lead to problematic situations in the real world with further consequences than merely 'feeling tricked' by the AI. To account for this, Doolhof could have included dialogue at the end of the game that tied the lessons learned back to a realistic scenario. Ultimately, further research has to be done in elements that could further push motivation to perform output review.

Despite these concerns, it should be noted that Doolhof was not fully ineffective at motivating participants, as 30% does report that Doolhof has motivated them to be more critical with their AI usage, indicating that there were elements present within Doolhof that achieved a motivational impact for a part of this sample. Furthermore, with 47% reporting neutral to this same statement, it should be kept in consideration that this question researched whether participants were motivated to be *more* critical with their usage by Doolhof. Considering the participant background - where all were already familiar with generative AI and potentially too with its limitations - it could simply indicate that these participants were already appropriately motivated to treat genAI critically, and therefore Doolhof did not further motivate them. Finally, it should be emphasized that even if Doolhof did not successfully motivate the majority of participants further, the majority of participants do indicate that Doolhof was clear in its embedded contents regarding the importance of output review, leading us to believe the first three steps of Bloom's Taxonomy (remember, understand and apply). were met.

The reflective results in figure 30a support this, with 8 students reporting that Doolhof raised their awareness and understanding about the importance of output review. This is still valuable in its own right, as a proper understanding of the importance of output review is the first step towards performing it.

5.3 Observational Data

In our observational data, we noted that male participants generally answered quicker but often incorrectly due to not sufficiently utilizing the game's tools to review the output and attempting to 'bruteforce' the puzzles. While not visible in table 1 (groups are randomly indexed), group 1 and 5 were those that finished first in our sample. Additionally, both these groups were also those with the lowest scores. In our design, we had intended for the score system to be a counter to this type of behavior. These results cast doubt on the effectiveness of this design, indicating that the point system by itself might not have provided enough motivation to perform output review. In contrast, the group that scored highest was often seen relying on output review during gameplay. Female participants generally seemed less inclined to bruteforce, on average having higher scores around the 500-600 mark. Additionally, it was observed that the group that had the highest score was very skilled at output review and understood the embedded learning content quite well. This does align with the score design: Incorrect answers were penalized more heavily than slow answers, to reward output review over bruteforcing.

In the previous section, we identified that the punishment for submitting wrong answers was possibly insignificant or too abstract, causing some player groups to not be as motivated, which could have influenced the score distribution. Furthermore, the game design itself also did not account for group play, as the point system was designed and tested on participants playing alone. In the actual experiment, players might have spent more time discussing with each other or trying various strategies of answers compared to a single person, which could have led to more mistakes being made or a larger time spent on a single puzzle as opposed to playing alone. The opposite is also true: Group play could have, in some cases, led to players progressing faster, with players filling in each other's gaps in knowledge and assisting one another. Overall, this could have led to more variance within the score distribution.

Two groups did not finish playing the game. One group was close, remaining in the last area when the 25-minute session was over and scores were collected. We consider the potential that this group overlaps with the participants who indicated being less familiar with games and therefore completed it slower - as this is the pattern we saw occurring in our playtesting group. This indicates the potential of us underestimating the time necessary for the study and that more time should have been allotted. Fortunately, this group received most of the embedded educational content and therefore provided a valuable perspective within the questionnaire. The other group did not make it past the introduction, nor did they inquire for help, and spent most of the session distracted by each other. While the teacher attending the experiment encouraged them to keep playing, Doolhof did not capture their attention over talking to each other. We believe external factors to be present here that are outside of the experiments' sphere of influence. This factor should be kept in mind for putting into perspective our results, as this group of participants is likely to have reported neutral for most Doolhof-related statements, thus slightly skewing the results.

5.4 Limitations

This subsection discusses limitations faced in this study. Firstly, our study had a relatively small sample size. 17 participants is hardly representative of the population (which would be all students aged 16-18 within the Netherlands). This means our results do not quite generalize to the population. While this should be kept in mind for future research, our results are still valuable and give us a worthwhile indication that this area of research has potential. Furthermore, since our sample was obtained through convenience sampling, sampling bias and selection bias could be present. In our sample, all participants were students of a teacher that is enthusiastic about the use of AI and is currently working on their own AI literacy curriculum. An effect of this can be observed within the background of our study: All participants were already familiar with generative AI. An underrepresented group thus exists, namely that of people who do not use generative AI technologies at all. This could have led to participants already being familiar with concepts such as output review. Consequently, those who already were familiar with output review could be more likely to mentally dismiss the embedded educational content, assuming they already know the contents that are being taught. Additionally, with the teacher being enthusiastic about AI and having attempted to give AI literacy education before, participants' opinions could be skewed positively in favor of generative AI technologies. An alternative theory is that this could have led to participants being more receptive to learning about output review, as it ties in to the teacher's existing curriculum.

Another relevant limitation was the unexpected technical difficulties we encountered during the empirical study. An unexpected problem was encountered with the schools' hardware blocking Doolhof, which was only found a short time before the empirical study took place. This problem was circumvented through allowing participants to play Doolhof on their personal hardware. This, however, caused a delay within the setup of our experiment. While students were still given the intended 25 minutes for gameplay, the time that students had to fill in the questionnaire was shorter than originally planned. Originally, 50 minutes were allocated in total for this experiment, of which 15 minutes were allocated for filling in the questionnaire. Since participants had another mandatory class after this experiment, this lack of time could have led to participants rushing to fill in their answers for the questionnaire. Consequently, this might have led to participants omitting detail during open questions (as it would be time-intensive to write it out in full detail). For example, participants could have intended to perform output review during hypothetical scenarios but might have neglected to mention that they would, due to rushing to fill in the questions. To slightly account for these technical difficulties, a portion of the questionnaires (3 to be exact) were instead manually submitted by students the next day, as the teacher allotted time for these students to finish the questionnaire during their class. This interestingly raises another concern for this subset, as the hours passing between the experiment and the questionnaire being filled in could also result in loss of detail in answers.

Further limitations were present in the design of Doolhof. With this thesis being limited in scope, it was determined that an educational game that could be played within a lesson hour (such as Doolhof) was more suitable to develop compared to a game that contained more hours of gameplay, which could be spread out over multiple lessons. The latter would require further development time, which was not feasible for this thesis. A larger development cycle could have also benefited Doolhof in other ways, such as further polishing to the puzzle design. Furthermore, it could be argued that

the design itself is limited in effectiveness. 25 minutes of exposure might be too short to have a lasting effect on remembering and motivation. Consequently, this choice of limiting the gameplay to 25 minutes might also have had a significant impact on our results. During our observations, we noted that two groups were not able to finish the game in this given time, meaning these groups did not explore the game's full contents. Furthermore, from the playtests it could be derived that those not familiar with games previously on average took longer than 25 minutes to finish the game, corroborating the theory that the groups who did not finish the games might be those with less familiarity with games. It should be noted that our playtesting group consisted primarily of people familiar with games. This could have influenced our design, skewing puzzle difficulty to be more difficult as it was tested on those familiar with these types of games.

An additional limitation within the design of Doolhof is the choice of how to represent Rosie. The Robot Companion was fake AI, not directly using generative AI for generating output. This limitation was caused by the various restrictions that utilizing real AI would have, such as using real AI directly in-game being quite computationally expensive and thus not meeting feasibility requirements. Further concerns are those of ethics: As this thesis aims to make clear: Generative AI is inconsistent and occasionally produces undesired output through hallucinations and other limitations. Therefore, using real AI without manually verifying the output could cause inappropriate content to be displayed to the targeted age group, or content that was otherwise unsuitable for the game experience. Developing a generative AI system that accounted for this was too resource and time-intensive for the purposes of this thesis. To account for these limitations, dialogue was instead manually scripted with the help of generative AI. However, this solution introduced bias in the design, with the designer deciding exactly how to modify and use the generated output for game dialogue, thus skewing how generative AI was represented to the player. Furthermore, when and how Rosie hallucinated was thus also influenced, as hallucinations in places where it would negatively affect the player experience were manually removed. However, it should be noted the hallucinations that Rosie displayed within the game were all real hallucinations that ChatGPT output during the prompting process.

Another noteworthy limitation is that the design of our empirical study merely regards the viability and feasibility of the requirements set in section 3.1.1. As stated, in the construction of these requirements, the educator was targeted as the end-user for desirability, while player concerns were addressed through feasibility and viability. Despite this, we have not evaluated desirability. The reasoning behind this choice is twofold: Our main interest within this research was finding how to effectively design a game that motivates players to perform output review. To evaluate this, a study focused on the player was strictly necessary. Furthermore, due to the aforementioned limited scope of this thesis, we were unable to arrange for a secondary study that regarded desirability. To evaluate how educators experience the use of Doolhof in the classroom, a much larger study would be required over a longer period of time, which was simply out of scope for the study. Through this limitation, one can debate the choice for choosing the educator as our end-user for desirability. Ultimately, this decision was made as we wished to develop an educational game that fit within the already existing efforts towards digital literacy and AI literacy within the Netherlands. For this purpose, our game needed to be suitable for classroom usage. Subsequently, this choice allowed us to build upon the existing competencies that the existing efforts regarding digital literacy intend to teach, such as computational thinking skills. This is reflected in the puzzle design, with certain

puzzles greatly benefiting from decomposition and pattern-seeking. Consequently, not evaluating whether Doolhof actually fulfilled its desirability requirements might have caused an oversight as to whether Doolhof truly fits within these existing efforts in digital literacy, as, for example, we have not managed to evaluate whether Doolhof appeals to both AI skeptics and enthusiasts.

Lastly, many of the methodologies we used involved self-reporting, which introduces various kinds of bias. For example, participants might have claimed to intend to perform output review because of some kind of social pressure, e.g., wanting to appear more competent or aware. Although self-reporting aligns with the goal of the research (as we mostly wish to investigate student perception of these tools and their intentions), different measurements could have been used.

6 Conclusion

In this thesis we explored how educational games can be used for the purposes of promoting output review. Firstly, in our related works, we dove further into why it is critical that we teach the importance of critically assessing generative AI output. We saw that hallucinations are a frequently occurring issue within generative AI, being a result of how AI is trained to learn patterns and predict outputs and as such, even the best AI hallucinate. Despite these limitations, generative AI has rapidly become adopted in various fields, with recent innovations such as within education, where it can be leveraged both by educators and students. Studies have shown that the use of AI tools within education can improve students' engagement and motivation, but that there are also limitations and pitfalls present. Using generative AI to substitute skills has been shown to hinder learning, and recent studies suggest the over-reliance on these tools might even go as far as to be harmful to learning. Furthermore, educators indicate other concerns within education, such as that of accidental plagiarism and students using generative AI to cheat. To alleviate the pitfalls of the rapid technological advancement and to prepare the new generations to be able to navigate this landscape, there have been efforts in the field of digital literacy and AI literacy, both locally and globally. These efforts also aim to teach people how to handle these new technologies in responsible ways, for example, through fostering a basic understanding on how AI technologies work. To expand on the goals set in these frameworks, this thesis introduced the term "output review", which we define as the act of analyzing generative AI output, intending to identify errors these technologies might make to alleviate the posed risks. Motivating people to perform output review is crucial for AI literacy, as it translates into critical AI usage. Furthermore, the nature of output review is relatively non-complex as a concept, making it a simple yet effective point to address after a basic understanding of how AI systems work has been achieved and is therefore an important concept to teach.

In conclusion, in this thesis we highlighted the importance of teaching output review. Furthermore, we successfully developed a game that taught people the importance of output review. We leveraged multiple theorems and frameworks to guide the design process, successfully embedding the learning goals within the game in a way that was preferred by students over traditional lesson methods. Additionally, the game made the importance of critically treating generative AI output clear. However, while Doolhof was effective at teaching the importance of critically assessing the output of generative AI, it had limited success in further motivating students to perform output review.

6.1 Future Work

Our research reveals potential for further research in this area. The aforementioned limitations in scope highlight various alternatives in our initial game design that remain underexplored in this study. One such possibility is the development of games that are designed to support educators over a longer part of their AI literacy curriculum. These games could have repeatable gameplay or more content not suited for a single gameplay session. These types of games could have various benefits, such as having a more lasting effect on motivation or improving knowledge recall for the embedded content. Furthermore, with these types of games allowing for longer exposure, it eliminates constraints from the design so that it allows for more abstraction of the embedded content in the design without that being counter to attaining the learning goal. Instead, this could allow for a wider possibility of narratives and design within the game that could better motivate participants. Similarly, further work could look at targeting different age groups. This would allow for the fostering of AI literacy skills at an earlier age. Additionally, it could also improve avenues for encouraging output review, as other age groups might be more receptive to certain types of motivation. Another avenue is that of group play. As noted during observations, participants seemed more engaged due to group play, and it was observed that group play was perceived as fun. We believe that group play offers an expansion in possibilities to engage players further that is a worthwhile area to investigate for future work.

Another previously mentioned limitation of this thesis was the focus on the player within the empirical study, which meant that the desirability of Doolhof (which held the educator as the user) was not properly analyzed. Future work regarding educational games to teach output review could instead target the educator, investigating their acceptance of Doolhof or similar games. Additionally, efforts could be made to ease adoption of games such as Doolhof within the classroom, by, for example, designing games targeted at developing the AI literacy of the educators instead of the students to support educators in keeping up with the rapidly developing landscape. Furthermore, there were other aspects of desirability left unconsidered. Doolhof did not build upon computational thinking skills or output review skills as much in the game design. While the assumption that computational thinking skills were present was certainly relied upon within the game design, Doolhof does not quite majorly expand on these skills, which neglects possibilities that the development of these skills could offer. Building on these skills could cause performing output review to be perceived as easier to perform, which could lead to an increase in motivation to use. Similarly, while Doolhof certainly clarifies the importance of output review and aims to motivate performing this, it teaches little about how to properly critically assess generative AI output. While this is not a major concern (as output review is still relatively non-complex in principle), building upon this through, for example, the use of the CRAAP test could have further benefits in encouraging the use of output review [22].

Our study made use of the DVF framework. This framework is, by origin, not intended for educational game design but to design commercial products. Further work is encouraged to investigate what other frameworks could be used, or alternatively, whether a framework could be developed that suits educational game design better. A different framework could also further take into consideration both player and educator desirability, as arguably better results can be achieved when a framework balances these concerns well. Additionally, research can be done towards using more theory-driven

guidelines for game design, such as Self-Determination Theory and Experiential learning theory [16]. Finally, this study marks a potential area of research in investigating what exactly motivates people to critically assess the output of generative AI. This thesis has illustrated that it is effective to leverage educational games as a method to teach importance of critically assessing generative AI output, however the results indicated that our designed game was not as effective for raising motivation. For future work, identifying which factors in the game design boosted motivation is a crucial factor in enhancing the development of effective educational games for motivating output review.

7 Acknowledgments

This thesis work was an immense project and could not have been completed without the help of many talented and helpful individuals. Firstly, I would like to thank Giulio Barbero and Dr. Anna van der Meulen, whose guidance was essential for this project. Furthermore, I extend my gratitude to the teacher and the school who made it possible to hold our study, who are not named for privacy reasons. This includes the participants who contributed to the empirical study.

Doolhof's development would not have gone as smoothly without the help of Ben, who gave a lot of technical advice regarding development and Irene González Manzano who contributed some custom art. Furthermore, I would like to thank all the folk who went out of their way to playtest Doolhof, of which in particular I would like to thank Yaell Brouwer, Jeroen Hartman, Alice Kramers, Mark Rengers and Michiel van der Bijl for their significant contributions through giving feedback.

Lastly, I would also like to thank those who proofread this very article. Significant contributions were made by Michiel van der Bijl, Jeanne Calon, Finlay Harper-Stevens, Sam van Hoijtema, Emily Jiang, Jim Kauffman, Jennifer Luo, Nicholas Pang, Mark Rengers, Lowen Steerneman and Davina Thomas.

This section is not exhaustive and many more people have contributed to this project than listed here. To those too, I extend my thanks.

7.1 AI Usage Disclosure

For both moral and legal reasons, we would like to clearly indicate which parts of this project have used (generative) Artificial Intelligence and which have not. No parts of this thesis document have been written with, or with the help of, generative AI. Furthermore, no code of Doolhof has been written by or with the help of generative AI. Generative Artificial Intelligence, specifically, GPT4-o[15] has been used for the purposes of generating some amount of the dialogue that the Robot Companion displays in Doolhof. However, all of this dialogue has been human-verified in order to ensure it is appropriate to display to minors, as well as processed and appropriately edited to ensure it fit within the context of the game.

[15]GPT-4o: https://platform.openai.com/docs/models/gpt-4o

References

[1] Abeer Alnuaim. "The Impact and Acceptance of Gamification by Learners in a Digital Literacy Course at the Undergraduate Level: Randomized Controlled Trial". In: *JMIR Serious Games* 12 (Aug. 2024), e52017. ISSN: 2291-9279. DOI: 10.2196/52017. URL: https://doi.org/10.2196/52017.

[2] Patrick Bassner, Eduard Frankford, and Stephan Krusche. "Iris: An AI-Driven Virtual Tutor for Computer Science Education". In: *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*. ITiCSE 2024. ACM, July 2024, pp. 394–400. DOI: 10.1145/3649217.3653543. URL: http://dx.doi.org/10.1145/3649217.3653543.

[3] Katrin Becker. "What's the Difference between Gamification, Serious Games, Educational Games, and Game-Based Learning?" In: *Academia Letters* (2021). DOI: 10.20935/AL209.

[4] Christian Bisson and John Luckner. "Fun in Learning: The Pedagogical Role of Fun in Adventure Education". In: *Journal of Experiential Education* 19.2 (1996), pp. 108–112. DOI: 10.1177/105382599601900208.

[5] Jeroen Bourgonjon et al. "Acceptance of game-based learning by secondary school teachers". In: *Computers & Education* 67 (Sept. 2013), pp. 21–35. DOI: 10.1016/j.compedu.2013.02.010.

[6] Aras Bozkurt. "Why Generative AI Literacy, Why Now and Why it Matters in the Educational Landscape? Kings, Queens and GenAI Dragons". In: *Open Praxis* 16 (Aug. 2024), pp. 283–290. DOI: 10.55982/openpraxis.16.3.739.

[7] Ting-Jui Chou and Chih-Chen Ting. "The Role of Flow Experience in Cyber-Game Addiction". In: *CyberPsychology & Behavior* 6.6 (2003), pp. 663–675. DOI: 10.1089/109493103322725469.

[8] Mihaly Csikszentmihalyi. *Flow: the psychology of optimal experience*. Jan. 2008. ISBN: 9780061548123.

[9] Fred Davis. *A Technology Acceptance Model for Empirically Testing New End-User Information Systems*. Jan. 1985.

[10] Fred Davis. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology". In: *MIS Quarterly* 13 (Sept. 1989), pp. 319–. DOI: 10.2307/249008.

[11] Sebastian Deterding et al. "From game design elements to gamefulness: defining "gamification"". In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. MindTrek '11. Tampere, Finland: Association for Computing Machinery, 2011, pp. 9–15. ISBN: 9781450308168. DOI: 10.1145/2181037.2181040. URL: https://doi.org/10.1145/2181037.2181040.

[12] Xiaoxue Du and Xi Wang. "Play by Design: Developing Artificial Intelligence Literacy through Game-based Learning". In: *Journal of Computer Science Research* 5 (Nov. 2023), pp. 1–12. DOI: 10.30564/jcsr.v5i4.5999.

[13] Yael Erez, Koby Mike, and Orit Hazzan. *Leveraging Computational Thinking in the era of Generative AI*. June 2024. URL: https://cacm.acm.org/blogcacm/leveraging-computational-thinking-in-the-era-of-generative-ai/ (visited on 07/15/2025).

[14] Michail Giannakos et al. “Games for Artificial Intelligence and Machine Learning Education: Review and Perspectives”. In: Sept. 2020, pp. 117–133. ISBN: 978-981-15-6746-9. DOI: 10.1007/978-981-15-6747-6_7.

[15] Amy Green et al. *Teen and Young Adult Perspectives on Generative AI*. 2024. URL: https://assets.hopelab.org/wp-content/uploads/2024/05/Teen-and-Young-Adult-Perspectives-on-Generative-AI.pdf.

[16] Christian Grund. *How Games and Game Elements Facilitate Learning and Motivation: A Literature Review*. Oct. 2015.

[17] Matti Haverila, Kai Haverila, and Caitlin McLauglin. “The Impact of Perceived Effectiveness of Non-Pharmaceutical Interventions (NPIs) on Attitude Toward Usage, Behavioral Intentions, and Actual Usage”. In: *SAGE Open* 14 (May 2024). DOI: 10.1177/21582440241253360.

[18] Ting Hsu and Tai-Ping Hsu. “Teaching AI with games: the impact of generative AI drawing on computational thinking skills”. In: *Education and Information Technologies* (May 2025), pp. 1–20. DOI: 10.1007/s10639-025-13624-3.

[19] Shih-Hua HUANG and Ting-Chia HSU. “Learning Effectiveness and Reflections on AI Literacy in Junior High School Students with Game-Based Learning and Problem-Based Learning”. In: *International Conference on Computers in Education* (Nov. 2024). DOI: 10.58459/icce.2024.4933.

[20] IDEO. *Design Thinking Defined*. URL: https://designthinking.ideo.com/ (visited on 07/15/2025).

[21] Jamie Jensen, Andrea Phillips, and Jace Briggs. “Beyond Bloom’s: Students’ Perception of Bloom’s Taxonomy and its Convolution with Cognitive Load”. In: *Journal of Psychological Research* 1 (May 2019). DOI: 10.30564/jpr.v1i1.421.

[22] Adeva Jane Kalidas, Esparrago-Kalidas, and International Journal Of Tesol Education. *The Effectiveness of CRAAP Test in Evaluating Credibility of Sources*. Jan. 2021. DOI: 10.11250/ijte.01.02.001.

[23] Stephanie Kastelein and Iris Verbruggen. *Functionele kerndoelen digitale geletterdheid*. May 2025. URL: https://www.slo.nl/publicaties/@24550/functionele-kerndoelen-digitale/ (visited on 07/29/2025).

[24] S.C. Kong and H. Abelson. *Computational Thinking Education in K-12: Artificial Intelligence Literacy and Physical Computing*. MIT Press, 2022. ISBN: 9780262543477. URL: https://books.google.nl/books?id=DrxNEAAAQBAJ.

[25] Nataliya Kosmyna et al. *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task*. 2025. arXiv: 2506.08872 [cs.AI]. URL: https://arxiv.org/abs/2506.08872.

[26] David Krathwohl. “A Revision of Bloom’s Taxonomy: An Overview”. In: *Theory Into Practice - THEORY PRACT* 41 (Nov. 2002), pp. 212–218. DOI: 10.1207/s15430421tip4104_2.

[27] Bill Kules. “Computational thinking is critical thinking: Connecting to university discourse, goals, and learning outcomes”. In: *Proceedings of the Association for Information Science and Technology* 53 (Dec. 2016), pp. 1–6. DOI: 10.1002/pra2.2016.14505301092.

[28] Georgios Lampropoulos. "Educational benefits of digital game-based learning: K-12 teachers' perspectives and attitudes". In: *Advances in Mobile Learning Educational Research* 3 (Aug. 2023), pp. 805–817. DOI: 10.25082/AMLER.2023.02.008.

[29] Richard Landers, Michael Armstrong, and Andrew Collmus. "How to Use Game Elements to Enhance Learning: Applications of the Theory of Gamified Learning". In: *Serious Games and Edutainment Applications: Volume II* (Mar. 2017), pp. 457–483. DOI: 10.1007/978-3-319-51645-5_21.

[30] Carnegie Learning. *The State of AI in Education 2025*. 2025. URL: https://discover.carnegielearning.com/hubfs/PDFs/Whitepaper%20and%20Guide%20PDFs/2025-AI-in-Ed-Report.pdf?hsLang=en.

[31] Marina Lepp and Joosep Kaimre. "Does generative AI help in learning programming: Students' perceptions, reported use and relation to performance". In: *Computers in Human Behavior Reports* 18 (2025), p. 100642. ISSN: 2451-9588. DOI: https://doi.org/10.1016/j.chbr.2025.100642. URL: https://www.sciencedirect.com/science/article/pii/S2451958825000570.

[32] Edwin Locke and Gary Latham. "A Theory of Goal Setting & Task Performance". In: *The Academy of Management Review* 16 (Apr. 1991). DOI: 10.2307/258875.

[33] Hairong Lu, Dimitri Van der Linden, and Arnold B. Bakker. "The neuroscientific basis of flow: Learning progress guides task engagement and cognitive control". In: *NeuroImage* 308 (2025), p. 121076. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2025.121076. URL: https://www.sciencedirect.com/science/article/pii/S1053811925000783.

[34] Pin Luarn, Chiao-Chieh Chen, and Yu-Ping Chiu. "The Influence of Gamification Elements in Educational Environments". In: *International Journal of Game-Based Learning* 13 (Jan. 2023), pp. 1–12. DOI: 10.4018/IJGBL.323446.

[35] Fengchun Miao, Kelly Shiohira, and Natalie Lao. *AI competency framework for students*. UNESCO, 2024. ISBN: 978-92-3-100709-5. DOI: https://doi.org/10.54675/JKJB9835.

[36] Iman Mirzadeh et al. *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*. 2024. arXiv: 2410.05229 [cs.LG]. URL: https://arxiv.org/abs/2410.05229.

[37] Amr M. Mohamed et al. "Empowering the Faculty of Education Students: Applying AI's Potential for Motivating and Enhancing Learning". In: *Innovative Higher Education* 50 (Oct. 2024), pp. 587–609. DOI: 10.1007/s10755-024-09747-z.

[38] Jared Moore et al. *Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers*. June 2025. DOI: 10.1145/3715275.3732039.

[39] Fiona Fui-Hoon Nah et al. *Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration*. 2023. DOI: 10.1080/15228053.2023.2233814. URL: https://doi.org/10.1080/15228053.2023.2233814.

[40] Davy Tsz Kit Ng et al. "AI Literacy: Definition, Teaching, Evaluation and Ethical Issues". In: *Proceedings of the Association for Information Science and Technology* 58 (Oct. 2021), pp. 504–509. DOI: 10.1002/pra2.487.

[41] Manuel Ninaus et al. “Acceptance of Game-Based Learning and Intrinsic Motivation as Predictors for Learning Success and Flow Experience”. In: *International Journal of Serious Games* 4 (Sept. 2017). DOI: 10.17083/ijsg.v4i3.176.

[42] Stichting Leerplan Ontwikkeling. *digitale geletterdheid*. 2024. URL: https://www.slo.nl/thema/meer/basisvaardigheden/digitale-geletterdheid/ (visited on 07/29/2025).

[43] OpenAI. *Introducing ChatGPT*. 2022. URL: https://openai.com/index/chatgpt/ (visited on 07/12/2025).

[44] Digitale Overheid. *What You Need to Know About AI Literacy*. 2025. URL: https://www.nldigitalgovernment.nl/featured-stories/what-you-need-to-know-about-ai-literacy/ (visited on 07/15/2025).

[45] ProductPlan. *What is MoSCoW Prioritization?* URL: https://www.productplan.com/glossary/moscow-prioritization/ (visited on 07/15/2025).

[46] Hannah Quay-de la Vallee and Maddy Dwyer. *Students' Use of Generative AI: The Threat of Hallucinations*. Dec. 2023. URL: https://cdt.org/insights/students-use-of-generative-ai-the-threat-of-hallucinations/.

[47] Rijksoverheid. *Digitale geletterdheid op school*. URL: https://www.rijksoverheid.nl/onderwerpen/digitalisering-onderwijs/digitale-geletterdheid-op-school (visited on 07/29/2025).

[48] Michael Sailer et al. “Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning”. In: *Learning and Instruction* 83 (Feb. 2023), p. 101620. DOI: 10.1016/j.learninstruc.2022.101620.

[49] Esther Shein. “The Impact of AI on Computer Science Education”. In: *Communications of the ACM* 67 (June 2024). DOI: 10.1145/3673428.

[50] Shubham Singh. *ChatGPT Statistics 2025 – DAU & MAU Data [Worldwide]*. June 2025. URL: https://www.demandsage.com/chatgpt-statistics/.

[51] Raymond Smullyan. *What is the name of this book?: The riddle of Dracula and other logical puzzles*. Prentice-Hall, 1978. ISBN: 0139550887.

[52] Lihui Sun, Zhen Guo, and Linlin Hu. “Educational games promote the development of students' computational thinking: a meta-analytic review Educational games promote the development of students' computational thinking: a meta-analytic review”. In: *Interactive Learning Environments* 31 (May 2021). DOI: 10.1080/10494820.2021.1931891.

[53] Tuan Sarifah Aini Syed Ahmad. “Application of the Bloom's Taxonomy in Online Instructional Games”. In: *International Journal of Academic Research in Business and Social Sciences* 7 (May 2017), p. 12. DOI: 10.6007/IJARBSS/v7-i4/2910.

[54] Timothy Teo. “Factors influencing teachers' intention to use technology: Model development and test”. In: *Computers & Education* 57.4 (2011), pp. 2432–2440. ISSN: 0360-1315. DOI: https://doi.org/10.1016/j.compedu.2011.06.008. URL: https://www.sciencedirect.com/science/article/pii/S0360131511001370.

[55] Gabriella Tisza. "The role of fun in learning". In: *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '21. Virtual Event, Austria: Association for Computing Machinery, 2021, pp. 391–393. ISBN: 9781450383561. DOI: 10.1145/3450337.3483513. URL: https://doi.org/10.1145/3450337.3483513.

[56] David Touretzky and Christina Gardner-McCune. "Artificial Intelligence Thinking in K–12". In: May 2022, pp. 153–180. ISBN: 9780262368971. DOI: 10.7551/mitpress/13375.003.0013.

[57] Jack Tsao. "Game-based Learning and Storytelling for Teaching and Learning". In: (2024). URL: https://commoncore.hku.hk/Teacher-materials/Guidebook%20on%20Game-based%20Learning%20and%20Storytelling.pdf.

[58] Jim Waldo and Soline Boussard. "GPTs and Hallucination: Why do large language models hallucinate?" In: *Queue* 22 (Sept. 2024), pp. 19–33. DOI: 10.1145/3688007.

[59] Jeannette Wing. "Computational Thinking". In: *Communications of the ACM* 49 (Mar. 2006), pp. 33–35. DOI: 10.1145/1118178.1118215.

[60] Ramazan Yilmaz and Fatma Gizem Karaoglan Yilmaz. "The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation". In: *Computers and Education: Artificial Intelligence* 4 (2023), p. 100147. ISSN: 2666-920X. DOI: https://doi.org/10.1016/j.caeai.2023.100147. URL: https://www.sciencedirect.com/science/article/pii/S2666920X23000267.

[61] Michal Zadik, Noa Bregman, and Nirit Soffer-Dudek. "Are You "In the Zone" Or "Disconnected"? An Investigation of Flow, Dissociative Absorption and Their Adaptive Versus Maladaptive Correlates." In: *Journal of Anomalous Experience and Cognition* 2 (Aug. 2022). DOI: 10.31156/jaex.23915.

[62] Jialing Zeng, Sophie Parks, and Junjie Shang. "To learn scientifically, effectively, and enjoyably: A review of educational games". In: *Human Behavior and Emerging Technologies* 2 (Apr. 2020). DOI: 10.1002/hbe2.188.

[63] Yue Zhang et al. "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models". In: *Computational Linguistics* (July 2025), pp. 1–45. DOI: 10.1162/coli.a.16.

[64] Wenting Zhao et al. *WildHallucinations: Evaluating Long-form Factuality in LLMs with Real-World Entity Queries*. 2024. arXiv: 2407.17468 [cs.CL]. URL: https://arxiv.org/abs/2407.17468.

A Characteristics

The Characteristics of our educational game. The first indented list describes the original requirements. Bullet points starting with • indicate characteristics.

Desirability (User Appeal)

1. The game **should** be suitable for classroom usage.
   - The game ideally fits within curricula that aim to promote AI literacy.
2. The game **should** avoid discouraging the use of generative AI overall, but merely encourage the use of output reviewing when AI systems are used, so that it remains specific.
   - We should avoid portraying artificial intelligence as harmful / unequivocally bad. The goal of this game is to encourage responsible usage, not discourage AI usage at all.
3. The game **could** teach tricks for good output review practices.
   - We could recommend the CRAAP or EVERY framework within the game.
4. The game **could** aim to build upon computational thinking skills that students aged 16-18 are expected to have.
   - Challenges within the game could rely on developing step-by-step solutions, pattern recognition, abstraction or decomposition of problems. However, this is not the game's main focus and we should avoid over-relying on the assumption that CT skills are present.

Feasibility (Technical Requirements)

1. The game **must** not be too short, as it will take away from the educative value if the learning content is too compressed.
   - In practical terms, the game must contain enough content and time to convey our learning goals we determined using Bloom's Taxonomy.
   - As a lower bound, we estimate the game needs to be a minimum 15 minutes.
2. The game **must** not be resource heavy, to ensure it is accessible even to those with older hardware.
   - Therefore, 3D characteristics would be strongly discouraged.
   - Furthermore, it would be best if the game's visuals are kept simple, to save on complexity.
3. The game **must** be completable for users not familiar with games.
   - A beta-test group that ensures this might be necessary.
   - An in-game hint system can be considered to ensure players do not get stuck.
4. Play sessions of the game **should** not be too lengthy, in order to ensure the user remains engaged.

- Balancing this with the Desirability, we wish for the game to work within traditional high school class hours. In the Netherlands, these are 50 minutes.
- Assuming set-up time for the game is required, and a post-game questionnaire is taken (both of which we average at 10 minutes). This yields us a window of 30 minutes as the outer bound on completion time.
- It should be noted the full game can be divided into several shorter sections that can be completed within this time frame, if the game itself is longer than a class.

5. The game **should** have a simple control scheme, so even users unfamiliar with games can easily navigate.
   - If our game requires character or button cursor movement, it is strongly advised we use the wasd and/or arrow keys for movement, as these form the most common control schemes. WASD keys are intuitive to those familiar with games, and the arrow keys are intuitive to those unfamiliar, as they provide easy affordances.
   - Likewise, for selection-wise control schemes, the mouse is commonly used. We can be inspired by old point-and-click games.

Viability (Practical Requirements)

1. The game **must** focus on educating the importance of critically reviewing the output of generative AI.
   - To achieve this, the game **should** utilize Bloom's Taxonomy in its design to ensure this focus is met.
   - The game must feature a story beat or gameplay feature that directly ties back to output reviewing generative AI.
   - This means that there needs to be something present within the game representing generative AI.
   - However, due to Feasibility requirements, this cannot be actual generative AI, as that would make it too resource-heavy. Furthermore, using AI might conflict with Viability as we would need to ensure that the AI we use is suitable for classroom usage.
   - This means that for such a feature, a form of fake AI has to be utilized.
2. The game **must** focus on motivating people to critically review the output of generative AI.
   - This means that the earlier mentioned story beat or gameplay feature needs to invite a critical mindset towards what we use to represent generative AI.
   - We can introduce rewards as a gameplay feature (for example, gaining points) to promote this critical mindset.
   - We can introduce punishment (such as losing points, or losing life) for neglecting to be critical of this feature or story beat.
3. The game **must** be intuitive to learn to play, even for those who do not regularly play games.

- To achieve this, the game **should** apply the principles of Goal-Setting Theory to its design, so that the player knows their objective.
  - In regard to interface design, we aim to follow the principles of HCI to ensure affordances are clear.
  - Goals in the game must be well-structured and feel achievable.

4. The game **should** aim to be engaging for the average high school students, taking in account that their attention span might be short.
   - To achieve this, the game **should** apply the principles of Flow Theory to its design.
     - Ensure the player is rewarded for completing tasks.
     - Goals in the game should not be overly complex.
     - Challenges in the game must be tested to ensure they are neither too easy or too difficult, so that the game remains engaging.
5. The game **should** directly tie back to critical thinking within the context of generative AI, as if it were too abstract about this connection, educative value might be lost.
   - Therefore, the story beat or gameplay feature introduced needs to represent generative AI directly so that there is no layer of abstraction present. The player must be aware of the nature of this story beat or gameplay feature.
6. Ideally, the game **could** clarify generative AI's stochastic nature and lack of consistent reliability.
   - Therefore, the story beat or gameplay feature introduced should convey generative AI's stochastic nature and lack of consistent reliability.

B Design Document

Below the original design document for Doolhof can be found. The structure of this document was inspired by the design document of the original Grand Theft Auto games[16]. Final design of elements in Doolhof might divert from the document presented below. In fact, it should be noted several ideas in this design document have been removed from the final game to avoid complexity.

B.1 Specification

B.1.1 Concept

This document specifies a design for the gameplay with the provisional title "Doolhof". It is designed by Alette Farzad for the purposes of their thesis project, which researches how educational games can aid in teaching students to critically review the output of generative AI.

Doolhof is a game in which the player finds themselves stuck in a maze. They are tasked with solving puzzles to escape. To aid them on this adventure, the player has access to a 'handbook' and

[16]GTA Design Document: http://gamedevs.org/uploads/grand-theft-auto.pdf

a 'robot assistant'[17] that can be relied upon to give them clues to the puzzles. The robot assistant may at times provide unreliable or unhelpful dialogue, to encourage the user to critically regard its statements.

B.1.2 Gameplay

Doolhof's gameplay consists of sets of puzzles interconnected through narrative gameplay; e.g. dialogue progression.

B.1.2.1 General Overview

Within this game there will be a variety of puzzles. The aim is for these puzzles to require complex thought to tempt the player into utilizing the available assisting tools. However, puzzles themselves should in theory be solvable without using these tools, as otherwise it might feel like an insurmountable task, which would decrease motivation.

We wish to simplify these puzzles as much as possible. In that regard, we decided to take inspiration from classical point-and-click games, eliminating the need to figure out a player character sprite and movement.

We wish for some of these puzzles to resemble day-to-day tasks that one might want to use artificial intelligence for. Furthermore, we aim to seek inspiration in existing educational games. Below, we describe the three different puzzle types that will be present in the game, and give an example of each.

- Missing Section of the Dice
  - A puzzle in which the player is shown images of dice from various sides. Then, the player is shown an image of a dice with one face blacked out. (perhaps this can even be done with a differently sided die?). The player is tasked to choose from a selection of options with their guess on which face needs to be filled in. This is inspired by the cuboid of the old ones, from the game Warhammer: Total War[18]
- Logic Puzzles such as Knights & Knaves
  - This puzzle will be the classic logic puzzle of there being one guard who always lies, one guard who always tells the truth. The player gets several dialogue options to choose from, however, is likely to need to rely on the actual source info from the handbook to solve this puzzle.
- Search-Query Puzzle
  - This puzzle is meant to reflect a real search query one might use generative AI for. The aim is to directly tie back the game to the actual context of generative AI.

[17] These names are tentative

[18] Cuboid of the Old Ones: https://steamcommunity.com/sharedfiles/filedetails/?id=1568246541

- An example of a puzzle of this sort might include a puzzle in which a question is posed that a player is unlikely to know the answer to, such as that of an obscure animal fact. (e.g. the latin name of an animal?). The player is prompted to type the answer out manually.
- The player may inquire for the robot assistant to give them an answer, however, the robot assistant's answer might not always be accurate. Alternatively, the player can navigate through an encyclopaedia that might be more intensive to navigate through, however they will gain the correct answer through this.
- The player can also ask the robot assistant how they can check this information.

B.1.2.2 Tools

In this game, the player has access to two tools that they can use to help them solve the puzzles, this in the form of the Adventurer Handbook and the Robot Companion. We briefly describe the gameplay of these objects below.
The adventurer's handbook is a book that can be accessed through the player user interface. We aim to make it functionally similar to a real book - holding an index, and page navigation. This adventurer's handbook will hold a random collection of information relevant to the puzzles in the maze, alongside irrelevant knowledge unrelated to the game, making it a task for the player to find the relevant content.

The robot companion is a display that can be accessed through the user interface. During certain puzzles or moments in the narrative, the robot companion will be highlighted, prompting the player to engage with it. The interaction robot companion will have three different states, based on the player's progression in the game.

During the first set of puzzles, the robot assistant will merely give a straightforward answer to the puzzles when asked for help. After a few instances, the player will be reminded that they can rely on the handbook to gather more accurate information.

During the second set of puzzles, after prompting the RC for help with a puzzle, the player may ask a follow-up question. The options for this consist of (1) asking the RC for their 'reasoning' for the answer), and (2) Asking the RC where they can find more information themselves about the topic, which will provide the player with an easy link to opening the guidebook on the right page.

At first, these options will be correct. However, during the third set of questions: The first option might answer that the guidebook claims that, even when it doesn't. Choosing the second option may lead to a non-existent source, or a wrong page (e.g. the right page exists in the book, but the RC gave the wrong page).

Outside of interacting with the robot during puzzles, there are instances where the player may freely interact with them, engaging in more casual conversation to introduce it as a character.

B.1.2.3 Other Gameplay Elements

This section describes a minor gameplay element we wish to include that might further enhance the player experience, namely the score system for brute-force prevention and feedback.

During the game design, it quickly became clear that we needed an element to encourage the player and give feedback, as well as have a mechanic that prevents the player from brute-forcing puzzles. We do this by keeping track of a score, displayed on the top of the screen. This score will be referred to in-game as 'showing how good you are performing as an adventurer', but otherwise holds little to no connection to the narrative. The player starts out with a standard score, and gains points based on how fast they answer puzzles (correctly). The player will be informed in the narrative that answering questions faster will contribute to a higher score. Incorrect answers will penalize the player by lowering their score. Aside from keeping track of a score, no leaderboards will be kept.

Furthermore, the player will be told through the narrative that other adventurers are also out for the treasure, and that time is of the essence. This will further encourage a sense of competition in the player, serving to encourage them to finish the puzzles in a timely manner, lest the player loses out on the treasure.

B.1.3 Story

The game is set in an abstract world that combines modern elements (such as a robot assistant) with a medieval setting, such as a classic maze/dungeon. The player character is a non-descript adventurer who throughout the game will remain unnamed and shall be referred to using neutral pronouns. They are a new adventurer that has headed out to the maze in search of treasure. With them they have an Adventurer's book that all adventurers have. They also have a Robot Companion, a new gadget they received from the adventurer's guild that uses artificial intelligence.

The robot companion is a tool carried by the player, which (is claimed) to use artificial intelligence. It has been given to them by the Adventurer's Guild (the entity who sponsors all adventurers). It holds a simple, helpful personality. It presents its statements unambiguously, even when it might be sharing incorrect information. We wish to create emotional engagement, showcasing the robot companion off as a 'friend'. It is intended to represent them in a way where it creates the impression that the player and the agent need to solve the puzzle together. The robot is trained on data of all adventurers.

An old adventurer appears in the story. They have rejected the use of the robot companion, and they save the player a few times upon selecting wrong choices. This can yield as a way to have wrong choices matter within the story.

The player's goal is to reach the end of the maze, where a treasure is said to be. Furthermore, the aim is to do so "before others reach the treasure", with the others in this instance being other adventurers that, while not present in-game, are claimed to be after the treasure as well.

The main theme of the narrative revolves around the critical use of generative AI. We aim to

present the robot companion as a useful tool that can definitely save time with its use, but can at times be wrong. We hope to clarify that output should not be relied upon blindly.

B.1.4 Visual Elements

The main theme of the narrative revolves around the critical use of generative AI. We aim to present the robot companion as a useful tool that can definitely save time with its use, but can at times be wrong. We hope to clarify that output should not be relied upon blindly. For our graphics, aim to go for a pixel-based art style, as many assets are available for this kind of visual effect. Furthermore, we believe it fits the vibe of old-school point-and-click games and mazes.

Figure 31: Basic User Interface Layout

Figure 31 shows the basic user interface design we have arrived at, incorporating the previously named gameplay elements. We also include a menu.

B.2 Technical Breakdown

For developing this software, we have chosen to use the open-source game engine Godot. There were several reasons behind this choice. First and foremost, Godot has plugins that make implementation of narrative systems rather trivial, such as Dialogic. Furthermore, Godot allows for the use of tilemaps, making designing the levels rather trivial as well.

For the purposes of development, we assume that the target systems maintain the following minimum specifications:

- 4 GB of RAM

- 1 GB of free storage space
- Processor: Minimum 1.1 GHz or faster, two cores.
- Operating System: Windows 10 version 10.0.19041 or higher
- Graphics: Forward+ renderer: Integrated graphics with full Vulkan 1.0 support

B.3 Gameplay Implementation

B.3.0.1 Dialogue

For implementing the narrative gameplay aspect we utilize Dialogic[19], a plugin for Godot made by Jowan Spooner. This plugin provides an interface that provides a highly adjustable dialogue system comparable to that of many visual novels. For the purposes of our project, this provides a rather simple and quick way to set up a dialogue to support the narrative, hence we have chosen to utilize it.

B.3.0.2 Player Tools

We implement the robot companion through utilizing built-in Dialogic functions. We construct them as a Dialogic character that can be accessed through the player UI, whose dialogue and timeline changes based on progression flags within the story.

The adventurer's handbook will be accessible through the player UI and can be accessed by clicking it at any time. The player will be able to navigate through the book by clicking buttons at the edges of the pages. Furthermore, the player will be able to skip to a specific page by utilizing the number keys. The book will contain a page index (indicating what page covers what topic).

B.3.0.3 Puzzles

Our puzzles will be implemented differently based on puzzle type. For Puzzle Type 1 and 2, we utilize a multiple-choice system implemented through Dialogic, where the user may select the answer using their mouse. For puzzle type 3, the player will be able to type using the computer's keyboard to give their answer.

B.3.0.4 Score & Statistics

We use a global script that can be accessed through signals to keep track of the player's score. Statistics will be kept track of in a global script that can be freely accessed, similar to the score. At the end of the gameplay, we generate a file locally to store the player statistics.

B.4 Acknowledgments

The structure of this document was inspired by the design document of the original Grand Theft Auto[20].

[19] Dialogic: https://github.com/dialogic-godot/dialogic
[20] GTA Design Document: http://gamedevs.org/uploads/grand-theft-auto.pdf

C Other Data

In this section, the graphs that were not included within section 4 are included. Notable is figure 35 which displays further statistics of our direct questions.

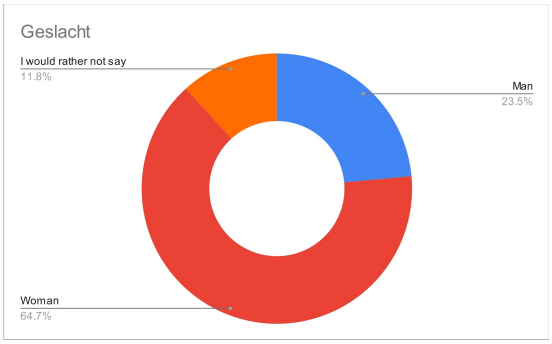


Figure 32: Gender Distribution of Study Participants

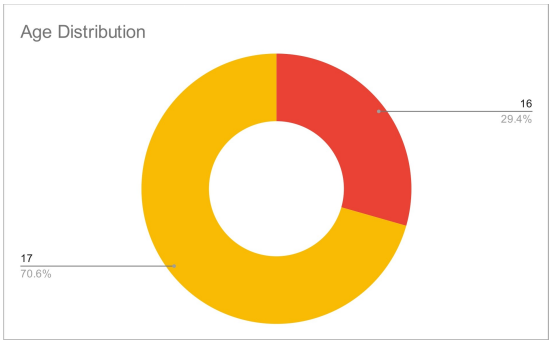


Figure 33: Age Distribution of Study Participants

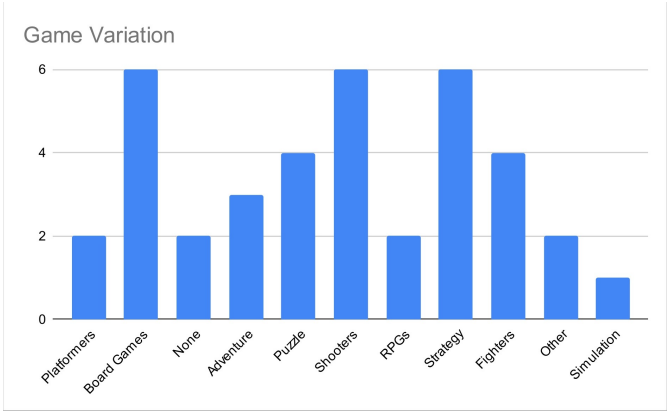


Figure 34: Game Variation Distribution of Study Participants

Catagory	Question	Mean	Median	Mode	STD. Dev
Perceived Ease of Use	The game was easy to control and navigate	3.294117647	3	3	0.8488746876
Perceived Ease of Use	The game was clear and easy to understand	3.529411765	3	3	0.9432422183
Perceived Ease of Use	I could play the game at my own pace and to my own wishes	4.294117647	4	5	0.7717436331
Perceived Ease of Use	The game gave me enough motivation and satisfaction	3.411764706	3	3	1.064120736
Perceived Ease of Use	The puzzles in the game were too difficult	2.823529412	3	2	1.131110854
Perceived Ease of Use	The handbook was difficult to use	2.823529412	3	5	1.629236559
Perceived Usefulness	I am learning better about the critical use of the output of AI systems with this game compared to other methods (such as books	3.176470588	3	4	0.8089572082
Perceived Usefulness	These kinds of games would be a good addition to our lesson curriculum	3.941176471	4	4	0.8993461677
Perceived Usefulness	I learn faster about how to use generative AI through this game	3.588235294	4	4	1.121317502
Perceived Usefulness	This game makes it easier to learn about generative AI compared to a normal lesson	3.235294118	3	3	1.147247345
Perceived Usefulness	I would rather learn about generative AI with this game then other methods (such as books)	3.705882353	4	5	1.159994929
Custom	The handbook gave me useful information	3.941176471	4	4	0.8269362306
Custom	Rosie was helpful	3.352941176	3	5	1.32009358
Custom	Rosie was not dependable	2.882352941	3	3	1.053704948
Output Review	I find it important to critically regard the information that generative AI gives me	3.764705882	4	4	1.032558217
Output Review	I will in the future output review the information generative AI gives me	3.529411765	3	3	0.9432422183
Output Review	I am aware that information generative AI produces is not always correct, and can be misleading	4.411764706	5	5	0.7122871199
Output Review	The game made it clear that it is important to treat generative AI more critically	3.529411765	4	4	0.7998161553
Output Review	The game motivated me to be more critical with my generative AI usage	3.058823529	3	3	0.9663454503

Figure 35: Mean, Median, Mode and Std. Dev of Direct Question answers

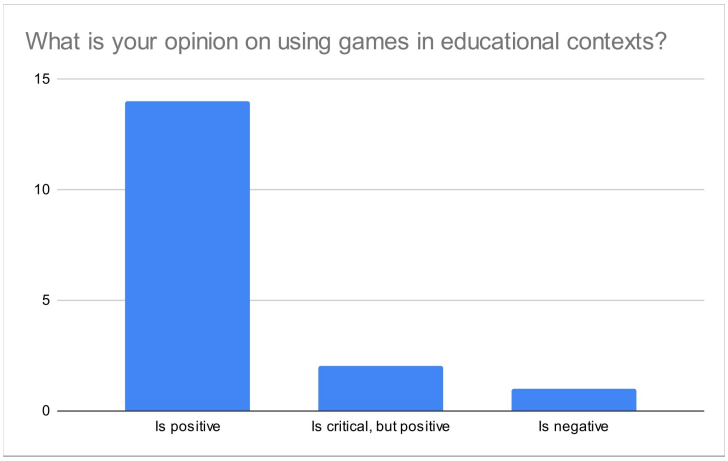


Figure 36: Participants' opinions on educational games

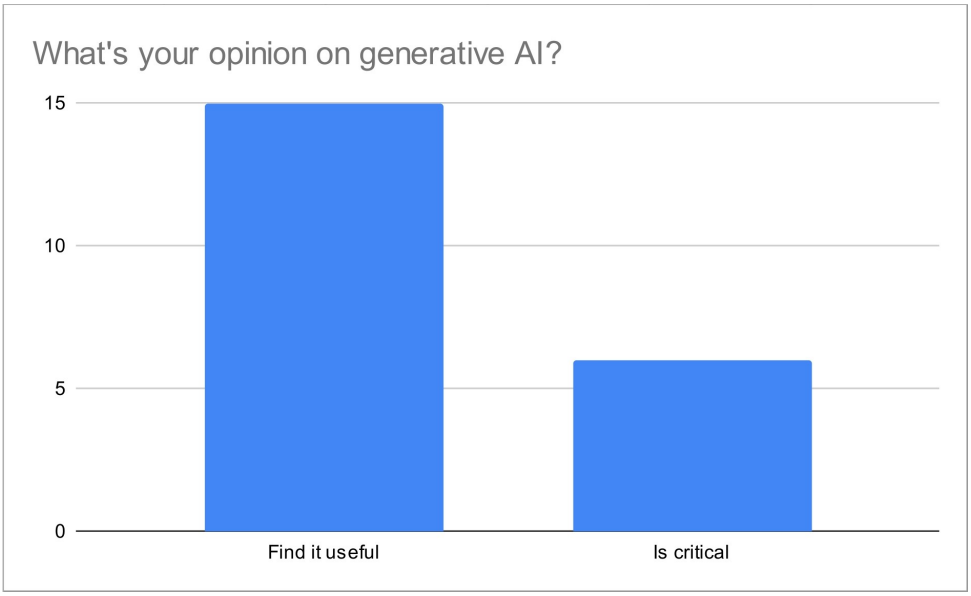


Figure 37: Participants' opinions on Generative AI

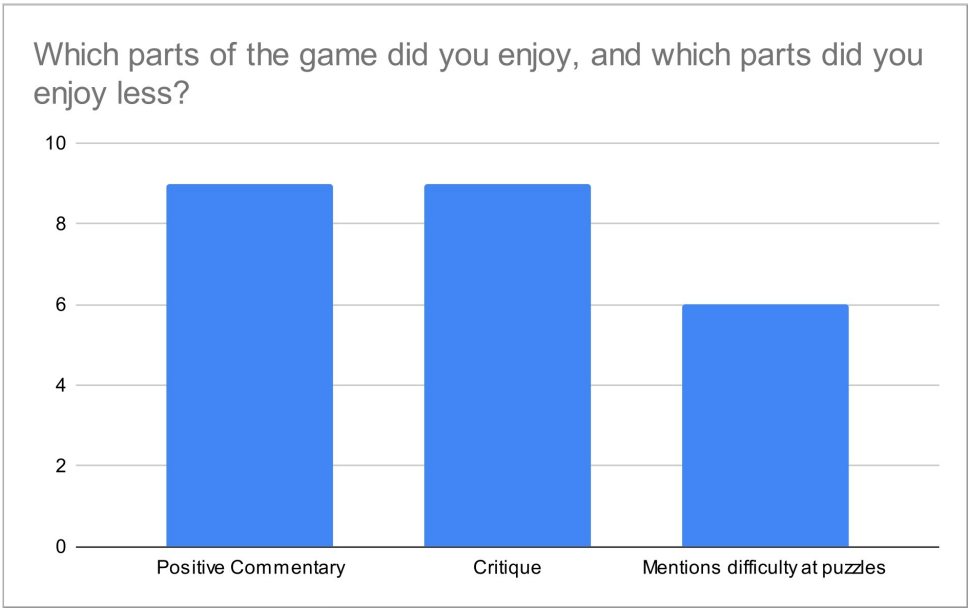


Figure 38: Participants' experience with Doolhof