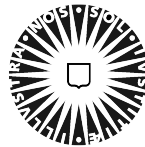


MASTER THESIS

Generative AI for Automatic Feedback Generation in Serious Games

Author:
Ruben BAGGEN

Supervisors:
Dr. Ivo SWARTJES
Dr. Julian FROMMEL
Prof. dr. Remco VELTKAMP



*A thesis submitted in fulfillment of the requirements
for the degree of Master*

in the

**Masters programme in Human-Computer Interaction
Department of Information and Computing Sciences**

August 15, 2024

UTRECHT UNIVERSITY

Abstract

Generative AI for Automatic Feedback Generation in Serious Games

by Ruben BAGGEN

This thesis investigates the use of Generative Artificial Intelligence (AI) for automatic feedback generation in serious games, focusing on identifying the characteristics of effective feedback in serious games, key design considerations, and practical implementation strategies. The narrative-based serious game "Take 5" is used for this study, employing an iterative design science methodology to develop and evaluate various prototype feedback systems. Multiple variants of an automatic feedback generation system designed around generative AI have been developed.

The study's iterative approach includes qualitative and quantitative evaluations with expert participants, where participants played and discussed output of the implemented systems, leading to insights that refine the feedback systems across multiple iterations.

The research identifies effective feedback as actionable, specific, personalized and motivational which are crucial elements for enhancing content in serious games. Key design considerations for integrating generative AI include leveraging contextual information about the player experience and characterizing goal, employing multi-prompt approaches for further consistency and relevance in the feedback provided, enhancing all identified content elements found.

The findings demonstrate that generative AI can improve feedback generation in serious games and that in 84% of the cases this feedback was preferable over the traditional already in-game feedback. This research contributes to the fields of serious games and educational technology by providing practical insights for implementing AI-driven feedback mechanisms in educational contexts.

Keywords: Personalized Feedback, Generative Artificial Intelligence, Serious Games, Game-Based Learning

Contents

Abstract	iii
1 Introduction	1
2 Related Work	4
2.1 Defining Game-Based Learning and Serious Games	4
2.2 Defining Adaptive and Personalized Feedback in Game-Based Learning	5
2.3 Defining Automatic Feedback Generation	7
2.4 Use of Generative AI in Automated Feedback Generation	8
3 Research Design and Methodology	11
3.1 Research Outline	11
3.2 Prototype Design	12
3.2.1 Context of the Game	12
3.2.2 Feature Specification	13
3.2.3 First Prototype and Iteration	14
3.3 Participants	15
3.4 Materials	15
3.5 Procedure	17
3.6 Data Preparation	18
3.7 Qualitative Analysis	18
3.7.1 Thematic Analysis	18
3.7.2 Comparison	19
3.8 Quantitative Analysis	19
3.8.1 Data Analysis	20
4 Results	21
4.1 Thematic Analysis	22
4.1.1 Data Preparation and Overview	22
4.1.2 Thematic Exploration	22
4.1.3 Thematic Framework	23
4.1.4 Generated Feedback Evaluation	24
Analysis 1	25
Analysis 2	25
Analysis 3	26
4.1.5 Parametric Feedback Comparison and Evaluation	26
4.2 Quantitative Analysis	28
4.2.1 Word Count Analysis	28
4.2.2 Sentiment Analysis	28
4.2.3 Personalization Measure Analysis	29
5 Discussion	31

5.1	Design Considerations	31
5.1.1	Prototype 1: Black-Box Approach	31
	Considerations	31
5.1.2	Prototype 2: Naive Approach	31
	Design	31
	Considerations	32
5.1.3	Prototype 3: Heuristic/Reasoning Approach	33
	Design	33
	Considerations	33
5.2	Research Findings	34
5.3	Limitations	36
5.4	Future Work	37
5.4.1	Recommendations from Participants	38
5.4.2	Future Research	39
6	Conclusion	41
A	Prototype Designs	43
A.1	Prototype 1: Black-Box Approach	43
A.1.1	Prompt:	44
A.1.2	System prompt:	44
	User prompt:	44
A.1.3	Generated Feedback:	44
	Participant 1.1	44
	Participant 1.2	44
	Participant 1.3	45
	Participant 1.4	45
	Participant 1.5	45
	Participant 1.6	46
	Participant 1.7	47
	Participant 1.8	47
	Participant 1.9	47
	Participant 1.10	48
A.2	Prototype 2: Naive Approach	49
A.2.1	Subtask 1: Feedback	50
A.2.2	System prompt:	50
A.2.3	User prompt:	50
A.2.4	Subtask 2: Example	50
A.2.5	System prompt:	50
A.2.6	User prompt:	51
A.2.7	Subtask 3: Tone	51
A.2.8	System prompt:	51
A.2.9	User prompt:	52
A.2.10	Subtask 4: Consistency	52
A.2.11	System prompt:	52
A.2.12	User prompt:	52
A.2.13	Generated Feedback	52
	Participant 2.1	52
	Participant 2.2	53
	Participant 2.3	53
	Participant 2.4	54

Participant 2.5	54
Participant 2.6	54
Participant 2.7	55
Participant 2.8	55
Participant 2.9	55
Participant 2.10	56
A.3 Prototype 3: Heuristic Approach	57
A.3.1 Subtask 1: Characterizing Goal	58
System prompt:	58
User prompt:	58
A.3.2 Subtask 2: Narrative Goal	58
System prompt:	58
User prompt:	58
A.3.3 Subtask 3: Example	59
System prompt:	59
User prompt:	59
A.3.4 Subtask 4: Feedback	59
System prompt:	59
User prompt:	60
A.3.5 Subtask 5: Concise	60
System prompt:	60
User prompt:	60
A.3.6 Generated Feedback	60
Participant 3.1	60
Participant 3.2	61
Participant 3.3	61
Participant 3.4	62
Participant 3.5	62
Participant 3.6	62
Participant 3.7	63
Participant 3.8	63
Participant 3.9	64
Participant 3.10	64
B Ethics Scan Utrecht University	65
C Informed Consent Participants	73
Bibliography	77

List of Figures

3.1	Model of Iterative Design and corresponding RQs	11
3.2	Landing Page	13
3.3	Intervention Introduction Page	13
3.4	Example team-member; Erika	14
3.5	Feedback Page	14
4.1	Word-Cloud of Initial Feedback Knowledge Analysis 1	23
4.2	Word-Cloud of Initial Feedback Knowledge Analysis 2	23
4.3	Word-Cloud of Initial Feedback Knowledge Analysis 3	23
4.4	Boxplots of Word Analysis per Iteration	29
4.5	Regression plot of F1 score Feedback and Script prototype 1	30
4.6	Regression plot of F1 score Feedback and Script prototype 2	30
4.7	Regression plot of F1 score Feedback and Script prototype 3	30
A.1	Design Prototype 1: Black-Box Approach	43
A.2	Design Prototype 2: Naive Approach	49
A.3	Design Prototype 3: Heuristic Approach	57

List of Tables

3.1	Overview of Demographics per Evaluation	15
4.1	Themes created for Thematic Analysis	24
4.2	Comparison Generated vs. Parametric Feedback and amount of Hal- lucinations in feedback	27
4.3	Sentiment Analysis of Feedback	29

Chapter 1

Introduction

In the rapidly evolving landscape of educational technology, automated feedback systems have emerged as pivotal tools in enhancing the learning experience. These systems, leveraging cutting-edge technologies, offer real-time, feedback to learners (Cavalcanti et al., 2021). Current applications typically involve comparing a student's response with a pre-defined correct answer, providing hints, corrections, or assessments that guide the learner toward the correct solution (Cavalcanti et al., 2021; Deeva et al., 2021).

Automated feedback systems are widely implemented in various educational contexts, such as programming, engineering, mathematics, and foreign language learning. These systems generate diverse types of feedback, including numerical scores, text-based comments, and visual aids, which collectively enhance the learning process by making it more personalized and efficient (Cavalcanti et al., 2021). Some systems are non-adaptive, offering the same feedback regardless of the learner's history, while others are task-adaptive or student-adaptive, considering individual learner characteristics and past performance. These systems are integrated into learning environments either as standalone applications or as plugins for existing platforms, facilitating their use in both online and blended learning settings (Deeva et al., 2021).

While the detection of errors has advanced, questions that center around generating in-depth learner feedback still need to be answered (Heift, 2022). Even though it is theoretically possible to write feedback for every possible combination of options or errors, it is also labor intensive. On top of that with current technology, it is impossible to account for unknown or potentially unlimited input, like open-ended questions, without sacrificing personalization (Shvets, Murtazin, and Piho, 2020).

The emergence of new large language models like generative pre-trained transformers, has the potential to revolutionize the generation of text-based feedback. GPTs offer a broad spectrum of applications and present significant opportunities for educational research, particularly in enhancing student engagement and aiding educators (Dwivedi et al., 2023).

The deployment of web-based pre-trained models like ChatGPT has sparked widespread interest, especially by educators in their application for automatic assessment and feedback systems (Kasneci et al., 2023b). Traditionally, many automatic feedback systems compare the student's solution to the desired outcome, and only a few implement more sophisticated Natural Language Processing techniques to generate feedback (Cavalcanti et al., 2021; Picca, Eberlé, and Jaccard, 2015).

Recent studies show that the transformer based large language model: ChatGPT is able to give effective feedback on English writing essays and assignments (Dai et

al., 2023; Escalante, Pack, and Barrett, 2023). This shows a promising future for the generation of feedback with this technology. In contrast, the paper by Kim, Shim, Shim, et al., 2023 states that ChatGPT is incapable of providing detailed feedback. ChatGPT is only able to evaluate linguistic knowledge and coherence in the written work. While it can offer basic feedback, it does not provide comprehensive evaluations that can drive learners to improve their communication skills.

On the other hand, ChatGPT has been speculated to have major impact and suggest that it is likely to offer significant gains in the banking, hospitality and tourism, and information technology industries, and enhance business activities, such as management and marketing (Deeva et al., 2021). It can be integrated into educational platforms to provide immersive, interactive learning experiences. Specifically, ChatGPT can be used to develop AI-powered educational games and simulations that offer hands-on learning opportunities, foster critical thinking, and enhance problem-solving skills and provide personalised content (Dwivedi et al., 2023).

Large language models are already successfully utilized in entertainment games to enhance interaction and immersion. They generate dynamic dialogue for non-player characters (NPCs), can also function as game masters in tabletop role-playing games, crafting narratives and responding to player actions in real-time (Gallotta et al., 2024).

Current research, however, lacks understanding of the full potential of LLMs as interactive player assistants. LLMs often struggle to accurately capture and interpret user intent, which is crucial for providing meaningful assistance in games. Future research is needed to develop methods to mitigate these issues, ensuring that LLMs can offer accurate, contextually appropriate hints, tips, and task management to enhance the player experience (Gallotta et al., 2024).

Furthermore, Large Language Models have significant potential to enhance procedural content generation in educational games. It is presumed to be able to generate rich, adaptive, and personalized content that tailors the learning experience to individual students. By training on existing educational content, LLMs can produce real-time text generation for various in-game elements like quizzes and tests, which adapt to the learner's needs (Amresh, 2023).

This approach not only reduces the development time and cost associated with content creation but, also promotes re-usability of systems (Westera et al., 2020; Amresh, 2023). AI has also been shown to make systems possible to account for exponential content growth, showing robust methods for systems that increase possible combinations (Blake et al., 2012). Still problems remain, serious games are not only limited to language learning and can be made for many different use cases (Susi, Johansson, and Backlund, 2007), and the feedback generated in these games comes in different modalities, timing and are used for different purposes (Yi et al., 2017).

Luckily, numerous advancements are rapidly evolving: ChatGPT allows for custom-made fine-tuned Chat-bots called GPTs (OpenAI, 2023), Google introducing their competitor to ChatGPT 4, called Gemini; promising even more multi-modality (DeepMind, 2023) and even an entirely different Neural Network, not based on transformers, promising to be 5 times faster than generative pre-trained transformers (Gu and Dao, 2023). In addition there are frameworks that try to overcome specific shortcomings like knowledge-intensive jobs (Lewis et al., 2020) or AI-Hallucination (Yin et al., 2023). With this rapid increase in research signs point to even more and faster

increasing advancements (García-Peñalvo and Vázquez-Ingelmo, 2023) opening up new doors for effective automatic feedback generation (Deeva et al., 2021).

This research is interested in gathering insights in effective methods of feedback generation utilizing this novel technology and evaluating its limitations. Solving the ever-growing need for serious games to further personalize the learning experience, promoting higher motivation, better learning outcomes, and player agency. All while reducing the time and effort to make serious games. For this the research is focused on solving the following design problem:

- **Improve** the effectiveness, efficiency, and scalability of text-based automatic feedback generation in serious games
- **By designing and evaluating** an automatic feedback generation system using generative AI
- **That satisfies requirements** for tailored feedback
- **In order to** help designers create more effective and tailored serious games, reduce design time, and support increased complexity

Following a design science methodology (Wieringa, 2014), the fundamental components of such a system are explored, developed, and evaluated using a previously developed narrative-based Serious Game called: "Take 5". This study follows an iterative design and development process aimed at creating an effective treatment to the problem. This research explored multiple designs and highlights different design approaches. To specify the goal further, and aim the development and evaluation process, the problem definition is broken down in multiple research questions.

Research Question 1: *What constitutes effective process-based feedback for automatic feedback generation in serious games?*

Research Question 2: *What are the key design considerations for integrating generative AI into serious games for automatic feedback generation?*

Research Question 3: *How can generative AI be utilized to develop an automatic feedback generation system within serious games?*

To answer these questions, an extensive amount of literature is reviewed to discover what serious games are and how they differ from and enhance normal learning, define what personalized and adaptive feedback is within these applications, and get an overview of automatic feedback generation, highlighting applications and potential concerns of new implementations. From this, a set of preliminary requirements is defined, that in turn acts as a starting point for the development of a specific implementation of a scoped-down automatic feedback generation system for the 'Take 5' game.

This system is repeatedly tested and reiterated until the feedback provided is deemed sufficient by expert evaluators and the researcher. Results are evaluated and discussed, highlighting design decisions and lessons learned during the development. This work contributes to serious games and automatic feedback research in finding applicable solutions for future designers to implement.

Chapter 2

Related Work

2.1 Defining Game-Based Learning and Serious Games

Game-Based Learning (GBL), as characterized by Jan L. Plass and Kinzer, [2015](#), is an educational methodology that incorporates both digital and non-digital games to facilitate learning, engage students, and achieve specific educational objectives. In contrast, gamification refers to the integration of game elements into non-gaming contexts to enhance engagement or motivation. While game-based learning focuses on educating through gameplay using fully-developed games often termed serious games, gamification applies game mechanics to non-game environments (Krath, Schürmann, and Von Korflesch, [2021](#)).

Both game-based learning and gamification extend beyond traditional e-learning, improving learning outcomes in various ways (Krath, Schürmann, and Von Korflesch, [2021](#)). Serious games achieve this by having two goals: a player experience goal; the goal of the player during gameplay and the characterizing goal; the 'serious' goal, what the developers are trying to teach (Dörner et al., [2016](#)). These serious goals can be diverse, ranging from influencing effective and motivational outcomes, to behavior change, knowledge acquisition, motor skills development, enhancement of perceptual and cognitive skills, physiological outcomes, and the development of social or soft skills (Connolly et al., [2012](#)).

In addition to the characterizing goal, the player experience goals offer immersive environments where users can actively apply knowledge and skills, leveraging and benefiting from multi-modal interactions. These games effectively merge the capabilities of computer processing and data storage found in traditional Intelligent Tutoring Systems with a high level of user engagement and attractiveness (Bellotti, Berta, and De Gloria, [2010](#)). To achieve the player experience goals, game-based learning employs various design elements such as Role-Playing, Competition, Challenge, Scaffolding, and Immediate Feedback (Qian and Clark, [2016](#)).

The theoretical foundation of these design elements has been extensively researched in academia. These elements draw from theories in sociological and cognitive psychology, such as self-determination theory, flow theory, experiential learning theory, and constructivist learning theory (Krath, Schürmann, and Von Korflesch, [2021](#)). The theories inform the scaffolding of games, enabling them to adapt challenges to the learner's level, personalize content to individual characteristics, and drive practices for more effective and constructive feedback (Krath, Schürmann, and Von Korflesch, [2021](#)).

Studies have shown that serious games can be as effective as conventional learning, with a notable preference among students for game-based methods (Hartt, Hosseini, and Mostafapour, 2020; Karakoç et al., 2022; Lei et al., 2022) and they are particularly effective in enhancing motivation and player engagement (Laamarti, Eid, and Saddik, 2014; Susi, Johannesson, and Backlund, 2007; Jan L. Plass and Kinzer, 2015). Additionally, the application of games in learning is versatile, spanning various domains from education to medicine, and even military training (Susi, Johannesson, and Backlund, 2007).

2.2 Defining Adaptive and Personalized Feedback in Game-Based Learning

In addition to the influence of game elements on learning, feedback is the most important part of learner interaction within games (Zeng, Parks, and Shang, 2020). Feedback plays a pivotal role in learning, exerting a substantial influence on the overall learning process. The purpose of feedback is to reduce discrepancies between current understandings/performance and a desired goal. However, improperly applied feedback can hinder learning significantly (Hattie and Timperley, 2007).

To help design this feedback Hattie and Timperley, 2007 devised three questions that need to be answered for effective feedback: Where am I going? (What are the goals?), How am I going? (What progress is being made toward the goal?), and Where to next? (What activities need to be undertaken to make better progress?). These questions correspond to notions of feed-up, feedback, and feed-forward.

The importance of feedback is reiterated in a study on feedback in serious games by Johnson, Bailey, and Van Buskirk, 2017. This paper describes critical characteristics that affect the effectiveness of feedback in serious games, including the content of feedback messages, the modality in which feedback is presented, the timing of feedback presentation, and learner characteristics. In addition, guidelines for each characteristic are provided and based in cognitive theory.

The content of feedback in Serious Games can be further broken up into outcome-based, also known as corrective, and process-based, otherwise known as explanatory feedback (Johnson, Bailey, and Van Buskirk, 2017) or elaborated feedback (Shute, 2008). Outcome-based feedback focuses on the end result of the player's actions, essentially indicating whether these actions were right or wrong. This approach is straightforward and aims at immediate correction, leaving the reflection on the answers to the user. On the other hand, process-based feedback is more comprehensive. It not only addresses the action's correctness but also the reasoning and strategy behind it, offering deeper insights and understanding. This reflection is crucial in educational contexts where the goal is not just to find the correct answer, but to understand the logic and reasoning behind it (Johnson et al., 2017).

The paper identifies two critical aspects of effective explanatory feedback: specificity and actionability. Specific feedback minimizes extraneous cognitive load by providing detailed insights into what is correct or incorrect, helping learners understand nuances and focus on areas needing improvement. For instance, it not only points out errors but also explains why an answer is wrong and suggests the correct approach. Actionable feedback further enhances this by offering clear, practical steps for improvement, such as hints, strategies, or step-by-step instructions, leading to higher learning outcomes and performance (Johnson et al., 2017).

The impact of explanatory feedback in educational games is explored in a study by Killingsworth, Clark, and Adams, 2015, comparing self-explanation (outcome-based) and explanatory feedback. Both approaches improved learning; no significant differences in effectiveness were observed. However, Yang, 2018 found that corrective feedback, which guides learners to correct responses or actions, enhances game-based learning, and motivates and engages players more. The study noted that students receiving regular feedback reviewed learning material more frequently, mitigating the drawbacks of non-corrective feedback. This further supports the importance of timing on feedback in serious games for both content-types, as described in Johnson, Bailey, and Van Buskirk, 2017.

Lastly, Johnson, Bailey, and Van Buskirk, 2017, delve deeper into considering diverse learner characteristics and feedback preferences. This importance of adapting to individual capabilities in e-learning is increasingly recognized (Shute and Towle, 2018). Adaptation in game-based learning, often referred to as scaffolding, customizes the educational journey for each player (Cai et al., 2022). Implementing adaptive content in a serious or educational game is however, harder than in entertainment games as it does not only need to work, but the learning goals also have to be met (Amresh, 2023).

Personalization in learning, defined as tailoring instruction to individual needs and preferences, is distinct from adaptation, which involves real-time instructional adjustment based on learner progress (Peng, Ma, and Spector, 2019). In game-based learning, personalized feedback not only enhances educational impact but also increases learner motivation (Erhel and Jamet, 2013). Research shows that personalizing content for serious game players fosters motivated usage and user acceptance (Streicher and Smeddinck, 2016), and further underscores the amplified efficacy of feedback, especially when also tailored to individual learner characteristics (Gavriushenko, Karilainen, and Kankaanranta, 2015; Streicher and Smeddinck, 2016).

However, such personalization often only extends to the sequence of interventions, not the content itself, leading to an overlap in 'personalized' and 'adaptive' learning terminology (Cai et al., 2022; Shvets, Murtazin, and Piho, 2020), especially in the context of automatic feedback generation. Deeva et al., 2021, found no clear distinction between feedback of systems that generate elaborated feedback, adaptive feedback, and personalized or individualized feedback. For the scope of this project, personalized feedback is defined as feedback that is tailored to the individual player's choices and behavior.

To summarize, the effectiveness of feedback in game-based learning, particularly serious games, hinges on its ability to be actionable, specific, personalized, and delivered at the right time and through the appropriate modality (Johnson, Bailey, and Van Buskirk, 2017). Actionable feedback provides clear, practical steps that learners can take to improve, thereby fostering a deeper understanding of the content. Specific feedback helps minimize cognitive load by pinpointing precise areas for improvement, making it easier for learners to focus their efforts (Johnson, Bailey, and Van Buskirk, 2017). While personalization makes sure the feedback is tailored to individual choices, promoting a better learning process.

2.3 Defining Automatic Feedback Generation

In digital learning, unlike traditional settings, there is often no instructor present to provide immediate and constructive feedback. Instead, feedback mechanisms rely on automatically generated responses, which can be implemented in various ways, including metrics comparison, Natural Language Processing, and Neural Networks (Cavalcanti et al., 2021). While these technologies have advanced in assessing learner responses, they tend to offer standardized feedback, lacking personalization and depth (Heift, 2022).

This practice of generating automatic responses is part of Natural Language Generation. As defined by Reiter and Dale, 1997: Natural Language Generation (NLG) is a field of artificial intelligence and computational linguistics focused on creating computer systems that generate understandable human language texts from non-linguistic data. It utilizes a series of algorithms specifically Natural Language Processing to generate; summaries, explanations, reasoning, or answers (Reiter and Dale, 1997).

Within natural language generation three approaches exist: modular, planning, and integrated/global. The modular architectures involve distinct divisions among natural language processing sub-tasks, with each task being clearly separated and building upon each other. The planning approach views text generation as a planning process. It is linked to a long tradition in AI of a problem-solving task and offers a more integrated, less modular, more interconnected, perspective on the various sub-tasks of natural language generation (Gatt and Kraemer, 2018).

The main used natural language generation practices in serious games are natural language processing algorithms, with current applications applying the older; modular, approaches (Picca, Eberlé, and Jaccard, 2015). In the study by Picca, Eberlé, and Jaccard, 2015 an overview of natural language processing technologies used in Serious Games is given. Natural language processing is highlighted as a tool for enhancing user interaction and learning, allowing for a more engaging and personalized experience. This technology enables serious games to adapt dynamically to individual player needs, tailoring content and difficulty levels to improve learning outcomes, with automatic assessment and feedback generation.

The use of modular natural language processing in serious games can be seen in various implementations, such as "W-Pal," "I-Star," and specific military training games. For example, "W-Pal" and "I-Star" (Johnson et al., 2017) are designed for computer-assisted language learning (CALL). These systems integrate game elements with educational content, providing a platform for language learners to practice and improve their skills through interactive scenarios. Similarly, in military training, games like those described in "Pedagogically structured game-based training" (Hill et al., 2006) utilize modular NLP to offer dynamic and context-sensitive feedback, enhancing the realism and effectiveness of training exercises.

However, currently the dominant trend in natural language generation is integrated or global approaches. These approaches do not strictly separate tasks but instead interweave them, often heavily relying on statistical learning methods such as machine learning to bridge between non-linguistic inputs and linguistic output (Gatt and Kraemer, 2018). This is done by training a Neural Network on existing content to produce more content of the same type and style (Yannakakis and Togelius, 2018).

Integrated generation methods can further be categorised, as found in the research by Deeva et al., 2021 evaluating multiple methods for a feedback generation system. As found in the paper there are three groups: data-driven feedback engines, expert-Driven feedback engines, and mixed feedback engines.

Data-driven feedback engines use student data to derive rules for feedback delivery through data mining techniques like ML and NLP. Expert-driven feedback engines rely on rules set by educational experts or theories combined with, a simple rule-based algorithm to decide which feedback is best for the student. Mixed feedback engines combine expert knowledge with insights derived from student data. These systems benefit from the blend of expertly crafted rules and insights derived from data analysis to provide more nuanced and contextually relevant feedback.

Data-driven systems are often more adaptable and personalized but require robust data and sophisticated technology. Expert-driven systems are more predictable and can be more reliable in certain domains but may lack the adaptability and personalization of data-driven approaches. In many cases, a mixed approach that combines the strengths of both data-driven and expert-driven methodologies offers the most effective solution (Deeva et al., 2021).

2.4 Use of Generative AI in Automated Feedback Generation

Even though, as Heift, 2022 research into Intelligent Computer Assisted Language Learning concluded that current integrated models are still lacking in generating in-depth learner feedback and are instead only providing general non actionable feedback, transformer models provide a way to solve this (Dai et al., 2023). Introduced in 2017, transformers redefined the use of machine learning for natural language generation by focusing on attention mechanisms rather than recurrent or convolutional structures. Its key elements include self-attention mechanisms and an encoder-decoder neural architecture. This architecture efficiently captures relationships within sequences, enabling parallel computation and better handling of long-range dependencies.

However, Neural Networks can have different structures and optimizations to achieve different goals like; data augmentation, text-to-text and text-to-image generation, generating all kinds of content like images, text, sounds, and 3D models (Gozalo-Brizuela and Garrido-Merchan, 2023). As stated, one of the drawbacks of Data-Driven Feedback engines is that they needed a lot of prepared and specific data to generate good-quality output (Deeva et al., 2021). Conversational Interfaces implementing large language models like: "ChatGPT", that are already pre-trained on big datasets make the use of machine learning more accessible and therefore mitigate a big part of this problem (Hassani and Silva, 2023). Because of this easy access, the arrival of the latest text-to-text generation model, the AI and machine learning field has gained a lot of public interest (García-Peñalvo and Vázquez-Ingelmo, 2023).

With the release of new large language models to the wider public, research into the generation of artifacts has been given one umbrella term: generative AI (García-Peñalvo and Vázquez-Ingelmo, 2023). To better understand what is meant by this novel term, García-Peñalvo and Vázquez-Ingelmo, 2023 researched and defined generative AI as an AI-driven process that involves training models with input data to generate new content by learning patterns within the available data and creating outputs that represent these patterns in new ways.

ChatGPT's ability to personalize learning and support has been found to be a notable advancement across literature; customizing educational experiences and providing personalized feedback, particularly in language skills development, addresses individual learner needs effectively (Javaid et al., 2023; Kasneci et al., 2023b; Rasul et al., 2023; Baskara, 2023). Similarly, Rasul et al., 2023 and Mørch and Andersen, 2023 underscore ChatGPT's role in facilitating adaptive learning by offering tailored educational experiences, promoting student agency and creativity. Moreover, Liu et al., 2023 and Owan et al., 2023 highlight its application in evaluating problem-solving abilities and automating personalized assessments, suggesting potential transformations in traditional assessment methods.

On top of better assessment language learning, like GPTs have significantly enhanced feedback generation in writing skills by providing detailed, student-specific feedback. For instance, ChatGPT has been employed to improve English writing skills, offering tailored guidance that aligns with individual learning styles (Escalante, Pack, and Barrett, 2023). Similarly, in Chinese language education, GPTs have supported learners from low-income backgrounds, providing personalized assistance that adapts to diverse needs, though this approach still relies heavily on student input (Li, Li, and Cho, 2023). ChatGPT has also been explored as a tool to support teacher feedback in English writing, complementing traditional teaching methods with AI-generated insights (Guo and Wang, 2023).

In the realm of programming education, GPTs have proven their ability to handle complex, subject-specific queries. AI-driven systems, such as those employing ChatGPT, offer adaptive feedback for computer science assignments, surpassing traditional methods in flexibility and precision (Nysom, 2023). The integration of large language models in web development resources has further enhanced the learning process by providing diverse and context-specific code explanations (MacNeil et al., 2023). In addition the study by Dai et al., 2023 investigates ChatGPT's efficacy in generating process-focused feedback, marking a shift toward more nuanced and context-aware educational tools. Highlighting deeper, more analytical feedback styles in data science education.

AI's versatility extends to specialized educational tasks as well. In training scenarios, fine-tuned models have been used to enhance questioning skills through interactive chatbot experiences for training for child therapy sessions (Røed et al., 2023).

Further, extending the feedback process by reviewing multiple data types, the system for Automatic Evaluation of Machine Generated Feedback for Text and Image Data (Goyal et al., 2022) employs a synthesis pre-trained model to handle multimodal inputs, reflecting the increasing multimedia nature of modern education. In addition the system developed by Liu et al., 2020 also stands out for its ability to integrate and process multimodal data, offering tailored feedback that addresses the unique learning needs and styles of individual students.

No papers were found that implement this novel technology for feedback generation in the serious gaming setting. However, from both a multidisciplinary and an education perspective, automatic feedback generation in serious games is highlighted as a transformative implementation of generating content that could benefit from recent developments (Dwivedi et al., 2023; Kasneci et al., 2023b).

Integrating ChatGPT into educational feedback systems brings challenges and concerns. Primary issues include model performance and accuracy, with apprehensions

about their effectiveness in processing educational content, raising concerns over inaccuracies and limitations in offering deep insights (Yan et al., 2023; Gamage, Dehideniya, Xu, et al., 2023). Balancing the transformative capabilities of these technologies against the associated risks determines the future trajectory of AI in education (Yan et al., 2023).

The exploration of LLMs in educational settings has highlighted their potential to revolutionize feedback mechanisms in serious games (Dwivedi et al., 2023). These models can offer personalized, process-focused feedback, enhancing the learning experience by tailoring content and guidance to individual preferences and progress. This capability is particularly valuable in areas like language skills development, where nuanced and specific feedback can significantly impact learning outcomes (Escalante, Pack, and Barrett, 2023).

While the integration of ChatGPT and similar models into serious games is still in its early stages, the potential for these technologies to transform automatic feedback generation is substantial (Dwivedi et al., 2023). They offer the possibility of more interactive and personalized learning experiences, aligning with the goals of adaptive and personalized feedback systems. However, challenges such as ensuring accuracy and managing potential biases must be addressed to fully realize the benefits of these advanced technologies in educational contexts (Yan et al., 2023).

Chapter 3

Research Design and Methodology

3.1 Research Outline

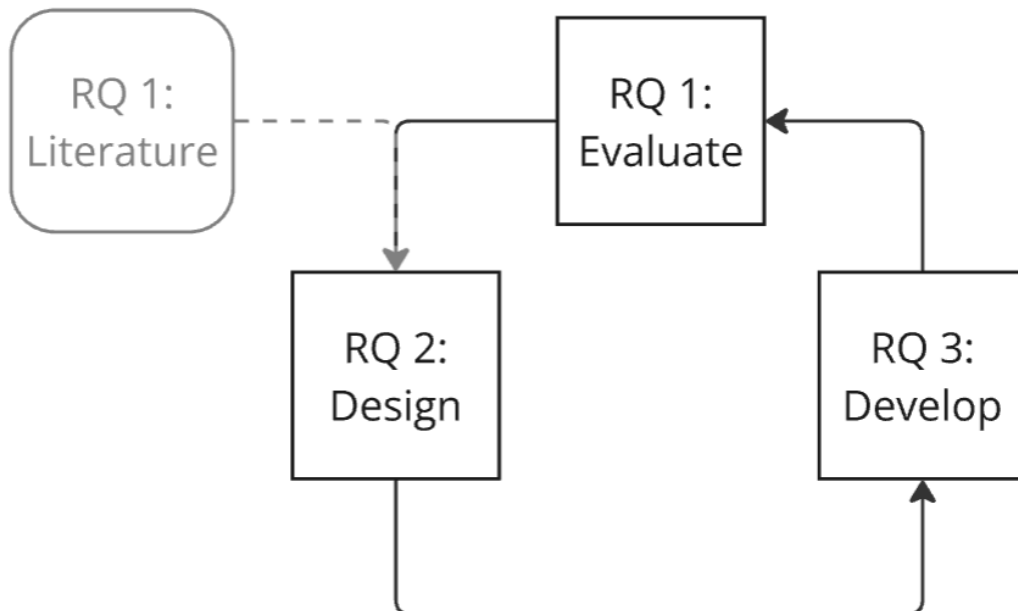


FIGURE 3.1: Model of Iterative Design and corresponding RQs

This study employs an exploratory sequential mixed methods design, grounded in the design science framework as proposed by Wieringa, 2014. Its primary goal was to investigate how Generative Pre-trained Transformers (GPTs) can be utilized within serious games to enhance and personalize feedback. This study applies an iterative design in which multiple prototypes and evaluations were carried out in succession, to gather more insightful and applicable results for answering the three research questions. This chapter details the methodology and overall research design with its components.

The iterative process was repeated across three cycles to refine the approach and outcomes. Figure 3.1 presents a visual summary the iteration cycles and how the research questions are used to aid in the iterative process. The first iteration however is different from the later two iteration, as the design and development of this iteration happens prior to evaluation, and is informed by the findings in the related work chapter 2. To help understand how each iterations design process is carried out the

first prototype is described here in the method section. Later designs and design considerations based on results from the applied research methods are discussed in the Discussion chapter 5.

The evaluation of the system is further split in a qualitative analysis and a quantitative analysis. Both of these studies play a role in evaluating the novel developed prototype and gather more insights to answer research question 2. However, only the qualitative analysis is used to answer research question 1. Quantitative research is meant to gather deeper insight in aspects users cannot directly comment on and phenomenon across multiple participants. These insights further validate and confirm the efficiency of the prototype.

3.2 Prototype Design

Prototypes were developed as potential solutions to the design problem outlined in the introduction. Each prototype was integrated directly into the Take 5 game's source code and hosted by &ranj¹. The design choices for each prototype are categorized and informed by the findings in the problem investigation phase. The initial prototype defined here was based solely on literature, while subsequent iterations were refined using insights from the expert study described in phase 1. Details on the construction of the prototypes and the methodologies employed are discussed in this section, with the later designs and reasoning behind those presented in the Discussion chapter 5, as part of the reflection of the results found.

3.2.1 Context of the Game

As mentioned in the introduction, "Take 5" is a management training game developed by &ranj, aimed at teaching valuable soft skills to new managers or employees. This characterizing goal is based in training material created by &ranj. The game is built in Unity and available on both Android and iOS platforms.

"Take 5" includes 48 lessons, each lasting between 1.5 and 6 minutes, and is designed to be played in short bursts, such as during breaks, totaling approximately 182 minutes of gameplay. It covers seven key skills: powerful questioning, deep listening, giving and receiving feedback, building trust, being courageous, and being agile, with each intervention focusing on one or two skills. Players engage through five types of interactions—dialogue with forecasting, strategizing, audio conversations, keyword selection, and theoretical lessons, which are structured around Bloom's taxonomy levels (Krathwohl, 2002). The player experience goal of each episode is narrative-based and you need to help your 'colleagues' gain certain insights or help them feel better.

After completing each episode, players receive both outcome and process-based feedback, determined by their performance. A star rating, ranging from 1 to 5, is initially displayed. Corresponding process-based feedback is then provided, with distinct responses tailored to each star level. This standardized feedback, whether focusing holistically on the entire episode or on specific parts of the questions, is parameterized based on the episode's content and therefore referred to as parametric feedback in this study.

¹Serious Gaming Company in the Netherlands; website: www.ranj.com

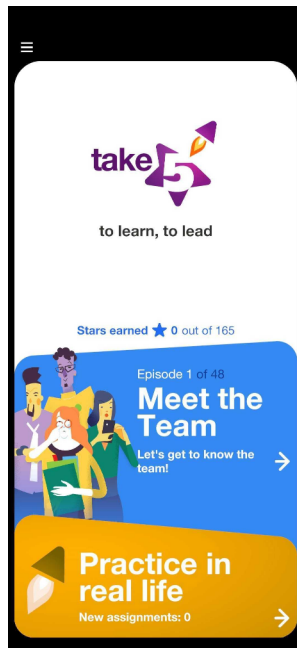


FIGURE 3.2: Landing Page



FIGURE 3.3: Intervention Introduction Page

To facilitate testing, a vertical slice of the game representing a single story arc was developed. This slice features a mix of interaction types and concentrates on two key topics: deep listening and powerful questioning. It includes six playable episodes and six theoretical episodes with varying levels of challenge. For this prototype, while the original outcome-based feedback in the form of a star-rating was retained, the process-based feedback has been replaced with a newly designed generated feedback feature specific to this research.

3.2.2 Feature Specification

The new feature intended to replace the existing in-game process-based feedback was developed in Unity, primarily using Unity's web request library to integrate with the GPT API developed by OpenAi. GPT API was selected due to its widespread availability and comprehensive API features and documentation at the onset of this research.

Accessible data and code from "Take 5" was fully utilized, However new data structures or ways to leverage data could be made, no restrictions were placed on code adaption. All capabilities of the GPT API were evaluated for this feature and utilized if needed. The aim was not to create a perfect solution for "Take 5" but to explore the challenges and prerequisites of developing a generative tool that can produce process-based feedback in serious games. The design of the feature emphasized generalizability and scalability, reflecting these priorities in the design choices made.

The GPT API has a simple structure that allows for HTTP requests in JSON format. Each request has multiple parameters. For this research only the: model, token window and temperature are utilized. Temperature was fixed on 70% for the whole duration of study. 70% is the base temperature that GPT API uses for its generation. In addition, only one specific model, gpt-4-0125-preview, was consistently used

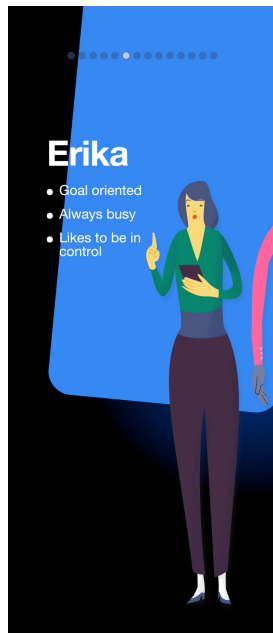


FIGURE 3.4: Example team-member; Erika

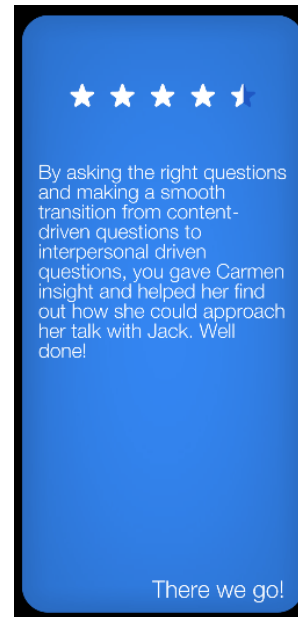


FIGURE 3.5: Feedback Page

throughout the research. This approach ensured that any observed improvements in feedback quality could be attributed to design changes rather than variations in the model or subsequent added parameters.

Additionally GPT API, features two main inputs; the system prompt and the user prompt. The system prompt acts primarily as the instructions of the API call. The user prompt is the input from the user and often kept more variable. Both are however posted in the request in one context window in the API. For this research it was chosen to utilize this function and keep the instructions of the feedback in the system prompt, and put the player data in the user prompt.

3.2.3 First Prototype and Iteration

To test out the full capabilities of the GPT API, the first iteration was a black-box approach, where a simple system was created without much reasoning added to the system, functioning as a starting point for this research. The subsequent prompting was also kept simple. The first iteration predominately focused on mapping the input and categorizing important aspects that the system would need to know.

The AI however is fed with existing data from the game and existing learning material to inform the generation, making this a mixed integrated feedback engine as described in literature. Furthermore a serious game has a characterizing goal and a player experience goal, working together to let players have the desired experience. Data is categorised based on these two aspects, where pink is used to denote the player experience goal and green for the characterizing goal as seen in figure A.1.

Data of the players choice is gathered after each input has been given and stored in a JSON file. Most of the data was already used by other functions in the game. However, the premise information is hard-coded in from external design documents provided by &ranj. The system is given a role as this has been shown to work very well for GPT prompts in scoping there task. No other specific prompting techniques

are utilized for the first prompt (Amatriain, 2024). The prompt for the first design can be found in appendix A.

Subsequent iterations were developed after evaluation of their previous design. Design considerations and design are presented as results in the Discussion 5 section and changes to this initial design and approach are discussed there. Iterations however all followed the basic structure of having both characterizing and player experience goal information, and worked only with the system and user prompt.

3.3 Participants

The study selectively recruited working professionals involved in the development, application, or study of serious games to evaluate the developed prototypes. These professionals, primarily sourced from the game developers of Take 5 and its associated Learning and Development consultancy, possess a more nuanced understanding of effective feedback within this domain and are able to give better pointers for improvement. Additional participants were purposively sampled from the researcher’s network, including those with (past) academic or professional expertise in (serious) games.

Furthermore, participants were allowed to volunteer through self-selection promoted at these organizations. Sampling continued in each new iteration until a quota of 10 participants per iteration was met, aligning with Creswell’s recommendations for qualitative research sample sizes ensuring meaningful exploration of diverse perspectives (Creswell and Creswell, 2017). Each participant engaged in only one iteration of the study, there is no overlap in participants between evaluations, totaling 30 unique participants across three evaluations.

Demographic details are summarized in Table 3.1. Participant ages ranged from 20 to 43, with an average age of 28.4. The gender distribution was nearly balanced, although the first iteration had a significantly higher number of male participants.

	Evaluation 1	Evaluation 2	Evaluation 3	Overall
Age Min-Max	20 - 33	23 - 43	23 - 36	20 - 43
Age Mean	26.6	32.1	26.5	28.4
Male	8	4	4	16
Female	2	6	6	14

TABLE 3.1: Overview of Demographics per Evaluation

3.4 Materials

The execution of this research necessitated the deployment of a variety of instruments, encompassing hardware, software, and metrics, all serving a role in the collection, analysis, and interpretation of data. Below, the specific requirements and functionalities of these instruments are found. In addition to the materials described below, before starting the research an ethics scan was performed by the university, results of this can be found in appendix B, furthermore an informed consent was provided to participants before participation, which can be found in appendix C.

Hardware:

- *Android smartphone:* The primary hardware requirement for participant engagement in the study is an android smartphone capable of installing and running the "Take 5" game application. If participants did not have a suitable device one was provided if possible
- *PC or Laptop:* Participants that partook in the research online needed a PC or Laptop with a working internet connection.

Software:

- *Unity:* Utilized for the development of the artifact and adaptation of the "Take 5" game. Unity serves as the foundational game development platform. This software enables the customization and integration of the feedback system evaluated and case study game in the study.
- *GPT API:* This generative model is employed to generate feedback based on participant decisions within the game. The API's advanced natural language processing capabilities and ease of use facilitate the creation of appropriate feedback. This is implemented at the heart of the automatic feedback generation system in this research.
- *Adapted version of Take 5:* A specially modified version of the "Take 5" game incorporating the automatic feedback system. This adaptation allows for feedback generation within the game's narrative and intervention frameworks, directly applied in the game.
- *Google Colab:* Used for data analysis, Google Colab provides a cloud-based environment for executing Python code. This tool supports the statistical analysis of data collected through the study, including player responses and feedback evaluation metrics.
- *NVivo:* NVivo is a qualitative data analysis (QDA) software designed to help researchers organize, analyze, and find insights in unstructured or qualitative data. Nvivo is used to code the gathered data from participants.
- *Miro:* Miro is an online collaborative whiteboard platform designed to facilitate visual collaboration. In the context of this research Miro is used to visually group codes together in groups and perform the thematic analysis.
- *JSONBin:* JSONbin is an online tool and service that provides a simple and convenient way to store and manage JSON data. It acts as a cloud-based JSON storage API that was integrated in the system to store user data from Unity in.
- *Zoom:* Zoom was used for online call appointments with iPhone users, as Zoom is the only free web-call application that allows both desktop remote access and recording at the same time.
- *Google Meet:* Google Meet was used for android users as this has a longer allowed recording time and does not kick participants automatically.
- *Qualtrics:* Survey hosting, distributing and analysis platform, used to create and store the informed consent form.

Models:

- *BERTScore*: BERTScore is a metric for evaluating the quality of text generation models, such as those used in natural language processing (NLP) tasks like machine translation, text summarization, and text generation.
- *Sentiment Analysis*: Sentiment analysis, also known as opinion mining, is the process of determining whether a piece of writing is positive, negative, or neutral. The Sentiment Score helps in quantifying this sentiment.
- *gpt-4-0125-preview*: For this research the latest model of the GPT API was chosen as a constant for feedback generation. At the time of the start of development this model was the latest text generation model.
- *Deepgram*: A platform specializing in automatic speech recognition (ASR) technology. It uses deep learning models to transcribe spoken language into text with high accuracy. Deepgram was chosen as it was one of the only models that can translate Dutch with high accuracy.

3.5 Procedure

Participants engaged in a two-stage qualitative study involving a vertical slice of the Take 5 game, featuring the built-in novel prototype. This condensed version showcased select episodes focusing on deep listening and powerful questioning skills.

After agreeing to partake, participants were offered the option to partake either in-person or online. Due to installation limitations on iPhones, users either used the researcher's phone or when online, remotely accessed the game on the researcher's laptop via remote access control in Zoom. After consenting and providing demographic information and self-assessed leadership skills (categorized as beginner, intermediate, or expert) in the informed consent, participants received a brief game orientation and extra background information on the game and narrative. Participants could speak both Dutch or English during the interview whichever they found most comfortable. The game was however played in English.

Before starting with the game-play, participants were asked if they wanted to play easier or harder episodes and the researcher selected for them the respective episodes. The story arc had 6 playable episodes with progressively harder episodes. Playing all episodes takes more than an hour. So to evaluate different levels of feedback and make sure participants had the ability to give feedback on the generated in-game feedback, the right difficulty needed to be found. As a starting point their self-assessed leadership-skills were used for this. However participants that reported having less or more experience with serious games or Take 5 could get differing starting episodes.

Furthermore, to get a better overview of both positive and negative score feedback, roles were also assigned, to either play optimally or poorly. Participants were asked if they would be comfortable with taking on the role as 'worst manager' and if they thought they would be able to evaluate this feedback.

Once participants knew their role and were to start, participants were asked one final question to evaluate their knowledge prior to play and to start the thinking process; "What would be effective feedback for you?". After this, they could play the game at their own pace. Participants were told to think-aloud while interacting

with the game. After each episode, participants received a star-rating and generated feedback, and were instructed to verbalize their thoughts about their received feedback. This was followed by targeted questions to delve deeper into their feedback experience.

Upon completing two to three episodes, the second comparative stage of the interview took place where participants were shown both the generated feedback they received to the in-game feedback with what they would have received in a standard version of the game, referred to here as the human-written feedback for easier recognition. They expressed preferences between the two feedback types, prompting further discussion based on their responses, on specifically what made them like or dislike either of the two.

After completing this comparison the research was concluded by asking the interviewee what they would like to add or how they see the future of this feature.

3.6 Data Preparation

Audio transcripts captured during the interview were transcribed using Deepgram AI a voice-to-text AI model capable of transcribing both English and Dutch. These transcripts were subsequently verified for accuracy and imported into the NVivo coding program for further analysis.

Further gameplay data was collected during expert interviews, which captured participants choices and corresponding generated feedback and parametric feedback. This data was stored in JSON format and sent to a remote JSONBin server after completing an episode. After each completed evaluation with 10 participants, the data was downloaded and cleaned by removing any entries from incomplete or empty episodes. On top of that the role the participant had during the gameplay was added as part of the script.

For further data analysis, cleaned JSON files were uploaded to Google Colab, a cloud-based Jupyter Notebook service that facilitates Python-based analyses. Within this environment, data was organized into arrays for scripts and generated feedback, with episode names serving as keys in a dictionary. This arrangement ensured that corresponding scripts and feedback were matched by their indices in the arrays.

3.7 Qualitative Analysis

The literature reveals that feedback is a complex domain characterized by diverse perspectives on what constitutes effective feedback. Research has established that the content of high-quality, process-based feedback should be actionable, specific, and personalized. However, the exact characteristics of such feedback and the methods to achieve it remain unclear. To further explore RQ 1: "What constitutes effective process-based feedback for automatic feedback generation in serious games?" a thematic analysis was carried out on the insights gathered from the expert participants.

3.7.1 Thematic Analysis

The analysis employed a thematic approach, following both deductive and inductive coding practices described by Braun and Clarke, 2006. Three requirements of effective feedback found in literature were used as deductive codes to further align

with the exploratory aims of this research and discover more requirements and how these take form. The thematic approach is further subject to inductive coding.

This method involved extensive coding of participant data to capture the nuanced expressions of participants accurately in Nvivo. Quotes from participants are summarized in specific statements referred to as codes. These were exported and pasted into Miro for a more visual attractive grouping process. Codes were color coded to highlight what they evaluated or answered. For instance, codes evaluating generated feedback were highlighted blue.

Codes were then grouped to identify recurring themes. At first groups were matched together with the deductive themes. If a code did not fit this group it was put in another pile together with similar codes. If a sufficient amount of codes relative to the amount of codes in other themes were grouped it was promoted to a new theme.

To further evaluate the prototype revealing further insights into the aspects of feedback that participants found useful or lacking, codes within a theme were further categorised as positive or negative. Insights from participants could then be used for further examination as to what the effect of the implemented prototype was on the generated feedback to help inform RQ 2.

3.7.2 Comparison

In addition to the thematic analysis highlighting important aspect of the generated feedback. The second part of the interview focused on comparing the already in game feedback participants would have received in the normal game against the newly generated feedback. This comparison was also carried out in a semi-structured interview style, further delve into the aspect participants voiced during the think-out-loud.

Opinions about the older parametric feedback were coded separately from the generated feedback and used for accessing the current state of the feedback and see where the generated feedback differs. In addition to this the overall sentiment of both was extracted by the researcher and used to assess which of the two participants found better.

Further analysis also included analysis to see whether or not the feedback aligned with each other, and if there is reason to believe there are hallucinations or untrue facts in the generated feedback. This was partially done by participants through questioning, but further done by the researcher if participants were incapable of spotting these discrepancies.

3.8 Quantitative Analysis

For gaining further insight about the generated feedback the study involved a quantitative analysis, focusing on the inputs and outputs generated by participants using the prototype. This analysis primarily aimed to validate the generated feedback received by participants and further support research question 2. No additional user analyses was required for this as the necessary data had been collected during the expert user evaluation in the interviews.

3.8.1 Data Analysis

Data from the scripts detailing player choices and the generated feedback was analyzed through three structured tests: Test 1, Test 2, and Test 3. These tests aimed to validate the consistency, accuracy, and impact of player decisions on the feedback, both within and across different iterations.

Test 1: Word Analysis Test This test analysed the consistency of the length of feedback generated in each iteration. A word count was performed for all generated feedback, and statistical analysis, including a boxplot representation, was used to assess the distribution of feedback lengths. The boxplot highlighted the upper and lower bounds, the mean, and the standard deviation for each iteration, facilitating a comparison of conciseness and variability in feedback length across iterations.

Test 2: Sentiment Analysis Test To evaluate the accuracy of the feedback's tone relative to the gameplay outcomes, sentiment analysis was performed using the transformers library on Google Colab. This analysis considered the intentional gameplay extremes; participants playing either deliberately well or poorly. The sentiment of the feedback was then compared against the player's role; a match in sentiment and role was deemed a positive result (i.e., positive sentiment for good performance and negative sentiment for poor performance). The proportion of correctly matched feedback was calculated for each iteration, providing a quantitative measure of tone accuracy across different iterations.

Test 3: Semantic Analysis Test This test measured the influence of player choices on the feedback generated. Using the BERTScore semantic similarity score, this analysis assessed the precision, recall, and F1-score between a candidate and reference text. Precision in this context reflected the proportion of relevant overlapping content captured in the candidate text, while recall indicated the extent to which the candidate text encompassed the pertinent details from the reference text. The F1-score provided a harmonic mean of precision and recall, balancing both aspects to gauge overall accuracy and comprehensiveness. This analysis was conducted for each script compared against other scripts from the same episode and for each related feedback compared against all other feedback for each episode, and for each a mean BERTScore was calculated. Scripts without comparative data within their episode were excluded.

Following the individual BERTScore evaluations, a multiple regression analysis was conducted to examine how script similarity influenced feedback similarity, considering the episode as an additional variable. This statistical approach allowed for a nuanced understanding of the relationships between player inputs and feedback outputs, on the assumption that very similar episodes should produce very similar feedback and very dissimilar episodes should also produce very dissimilar feedback.

Chapter 4

Results

This chapter presents the results of both the qualitative and quantitative analyses conducted on each prototype. Three distinct approaches were implemented: prototype 1 (the black-box approach, as detailed in the Methods section), prototype 2 (the naive approach), and prototype 3 (the heuristic approach). The final prototype successfully addressed the design problem, as it concluded in a solution that did not introduce any new significant issues. The outcomes that guided the development of Prototypes 2 and 3 are discussed in detail in Chapter 5, where additional insights into the design process and choices are also provided.

The results are presented collectively in each section, though they were obtained through a sequential, iterative process where each prototype built upon the insights and findings of the previous analyses. In summary, the following three prototypes were developed:

- **Black-Box Approach:** A single-prompt design that utilizes player choices, narrative, and theory to assess the baseline capabilities of the GPT API (figure A.1).
- **Naive Approach:** This approach addresses the limitations encountered in the black-box approach by refining and concisely structuring the feedback. It retains the same input as the first system but employs four prompts to generate feedback, breaking down key aspects into multiple steps (figure A.2).
- **Heuristic Approach:** Building on the naive approach, this method further refines the feedback generation process by better formatting the steps. It aims to capture the context of the "Take 5" game, enabling GPT to reason effectively and provide more targeted feedback (figure A.3).

The feedback generated by these systems is not presented as a result, as it does not constitute a direct outcome of the evaluation. However, to provide additional insight into the nature of participant feedback and to illustrate the changes across iterations, the generated feedback is included in Appendix A.

4.1 Thematic Analysis

4.1.1 Data Preparation and Overview

As described in the methods the thematic analysis was carried out per iteration and followed an inductive and deductive thematic analysis. Instead of using whole quotes, data was first coded in easier utterances or short conclusions for easier classification. Prior to evaluation duplicate or faulty codes were removed. Each evaluation was conducted with 10 participants totalling 30 participants across all evaluations. Mean Codes per participant for evaluation 1 were: 32.8, evaluation 2: 23.8 and evaluation 3: 22.2. 3 outliers were found across all iterations: two participant with sparse data and one with very rich data compared to the others in their iteration.

The first iteration in total had 294 codes. However, they were subdivided in codes exploring: themes of feedback (31 codes) prior to play, codes for evaluating the generated feedback (199 codes) and the parametric feedback (27 codes), and for additional topics that arose, like challenges participants faced during play (17 codes), like errors/typos and future work directions for either the app or the feedback system that are outside the scope of this research (20).

The second iteration totalled 239 codes with: 27 codes for themes in feedback evaluated before play, 170 codes about generated feedback, 18 codes for parametric feedback, 4 ideas for future work, and 14 problems with the app. Different from iteration 1, 4 codes voiced praise for the app.

The third and last iteration totalled 223 codes: 26 codes for feedback themes identified prior to play, 144 codes evaluating generated feedback, 22 codes for evaluating parametric feedback, 25 ideas for later and only 6 codes commenting on the game.

The thematic framework was built based on these codes. However while ideas for later and problems faced with the app during participation showcase valuable insight for the discussion chapter these results are not further addressed in this chapter or used for analysis.

4.1.2 Thematic Exploration

In order to get an initial understanding of what participants value in feedback, codes that were found in direct response to the introductory question: "What constitutes effective feedback for you (in serious games)?", were chosen to visualize using Word-Clouds. This is done to highlight prior knowledge of participants and gather initial thought to further inform the inductive thematic analysis process. This question is asked prior to gameplay and provides an overview of thoughts of participants going into the experiment. No further questions or examples were asked based on this.

The word-clouds for analysis 1 and 2 (figure 4.1 and figure 4.2) both have low word occurrences with the most occurring word only occurring 3 times. The last iteration (figure 4.3) has more occurring main words occurring 5 times. Stop-words, numbers and special characters are removed automatically, some remaining stop-words or unclear words that needed too much context to be clear, were filtered out by hand. In total, all analyses have less than 50 words.



FIGURE 4.1: Word-Cloud of Initial Feedback Knowledge Analysis 1

FIGURE 4.2: Word-Cloud of Initial Feedback Knowledge Analysis 2

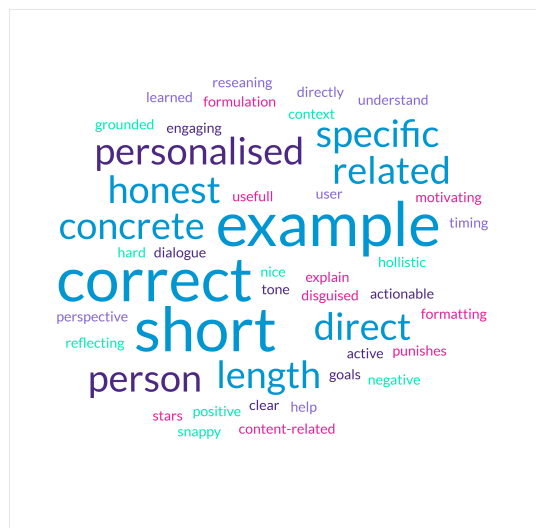


FIGURE 4.3: Word-Cloud of Initial Feedback Knowledge Analysis 3

4.1.3 Thematic Framework

New themes found were also used in subsequent iterations. To further test whether these fitted further inductive analysis on not categorised codes was carried out. However, for both thematic analyses two and three no further themes were created. This left some codes to not be categorised across all three analyses. These codes were either too general or too sparse to create new themes.

The feedback was initially structured around three deductive themes: **actionable**, **specific**, and **personalised**. These themes were established as high-level concepts essential for effective feedback. Subsequently, an inductive approach was employed to examine the feedback further, allowing for the emergence of additional themes

TABLE 4.1: Themes created for Thematic Analysis

Higher Level	Content Themes				
	Personalised	Specific	Actionable	Motivational	Other
Codes	82	77	60	27	77
Lower Level	Format Themes				
	Tone-of-Voice	Length	Example	Other	
Codes	83	46	97	60	

from the data. Codes that did not align with the initial themes were grouped into new categories based on similarities.

As seen in the previous section, data exploration word clouds highlighted several concepts frequently associated with effective feedback by participants. These included the **tone** of the feedback (positive or negative), the **use of examples**, and the **format**. These aspects, focusing primarily on the wording, language use, and structure of the feedback, suggested more granular thematic elements distinct from the initial high-level themes.

The analysis revealed that these format-related components significantly influenced the deductive themes. Notably, the presence of examples was found to enhance personalization, which in turn increased the specificity and actionability of the feedback. This interplay led to the development of a separate thematic layer within the framework, where such format elements were conceptualized as **format themes**. These themes serve as building blocks that support the overarching higher level **content themes**.

Three primary format themes were identified:

1. **Tone-of-voice:** This theme captures the spectrum from positive to negative feedback tones.
2. **Length of feedback:** This involves categorizing feedback based on word count, consistency, and redundancy.
3. **Examples:** This theme assesses whether feedback includes concrete examples that clarify and ground the content.

Additionally, one content theme related to the motivational effects of feedback in serious games emerged, emphasizing the impact of praise, compliments, and tonality on the player's experience. This theme is further referred to as **motivational** theme.

These newly identified themes were tested in subsequent iterations of the analysis to verify their fit and relevance. However, no new themes were generated in the second and third rounds of thematic analysis. As a result, some codes remained uncategorized across all analyses, being either too general or too sparse to form robust new themes.

4.1.4 Generated Feedback Evaluation

Codes that are categorised in the created thematic framework are further divided in positive or negative statements about that theme to better evaluate what participants liked and disliked about the generated feedback. This was used to better inform

design choices. In this section results are further explained and main points that were talked about by participants are summarized.

Analysis 1

The initial analysis revealed that out of 17 codes regarding actionability, only two were negative. This suggests that the initial prototype provided highly actionable feedback. The feedback was generally well-aligned with the relevant theories and narrative, with 13 codes perceived as positive and 7 as negative. The negative sentiments largely stemmed from participants feeling that some feedback was vague and lacked clear reasoning.

The feedback was personalized, with 21 codes highlighting positive aspects and 10 noting negative aspects. Participants appreciated the use of examples, which made the feedback feel tailored and impactful. Longer feedback was generally perceived as more personalized. However, some criticisms noted the overly positive tone and specific word choices, which sometimes made the feedback seem artificial. In terms of tone, while most feedback felt motivating, there was a consensus that negative feedback could be more strict, with only one code critiquing this aspect negatively.

One major overall issue identified was inconsistency, particularly in the length and use of examples. The feedback's length varied significantly, with some participants receiving feedback that did not fit on the screen. While longer feedback was acceptable if it provided sufficient detail and examples, there was a clear preference for consistent inclusion of examples across all feedback. Additionally, the quality of examples could be improved.

Other feedback indicated that the style resembled a "compliment sandwich" (positive-negative-positive format), which received mixed reactions. There were also suggestions for improved formatting, such as the use of paragraphs, bullet points, or bold text, to enhance clarity and readability.

Analysis 2

The second analysis was met with more negative opinions and across all themes except for motivational effect and length, was perceived as more negative. The theme of examples was met with the most negative codes (32 negative against 12 positive). Participants, while they always received examples, did not agree with the example, either because they thought it did not fit their choice or it was from a question that seemingly did not really impact the narrative, or was not explaining the theory very well.

This can also be seen in the specificity theme where 17 of the 25 codes were negative, predominantly because the feedback missed the overall bigger picture of what participants had to learn, or referred back to older theory that had nothing to do with the current episode participants played. The not fitting examples also had an effect on personalization, which was divided in 12 positive and 15 negative codes. The positive codes reaffirm the results in analysis 1 and reiterate that when receiving good fitting examples the feedback feels more personalised. However when receiving non fitting examples, or in this case also wrong fitting theory the feedback felt very general and not personalised.

Furthermore, this also had a big impact on actionability. The first iteration was met with great actionability reception. However in this prototype there were 6 negative

codes against 7 positive codes barely highlighting that the feedback was actually actionable. People stated that they did not agree with the feedback when receiving wrong information. Especially at lower star-ratings feedback was found to be more actionable. Highlighting that the the problem was due to the system choosing a right answer and presenting it as a wrong one.

The theme categorising opinions about tone reaffirms this, where from the 16 negative codes, participants highlighted that even when they got 5 stars the feedback was very negative and highlighted things they had to do better. Other remarks that were not directly linked to a theme, talked about that the feedback was incohesive and participants voiced their concern that they did not understand what they did wrong or what exactly was wrong about the feedback.

Analysis 3

The last analysis described in this process was again favorable. The feedback achieved good personalization because only one code from the 17 was found to highlight a negative aspect. This only stated that the play-style was reflected but the participant expected more. All other codes in this theme highlight that the play-style is reflected in the feedback and they recognise their choices.

Actionable was also good with 11 codes of participants agreeing that the feedback gave clear actionable steps that are even applicable outside the game. 3 codes were found to describe that feedback was lacking and the participant did not really understand what they should have done differently. This sentiment can also be seen in the specificity theme where the split is equal 7 for both negative and positive codes. Participants agreed that the theory and narrative is reflected in the feedback. However, they especially address their concern that the theory was not linked that in-depth to their example.

The reason for this can be seen in the example theme, which together with the length and tone themes were all found to have more negative codes. Participants got an example in all cases. However, especially for participants that achieved 4.5 stars, the answers were not highlighted enough. Participants expected to be shown the exact answer they were missing. In this prototype the example and theory were also linked more and the correct answers were given as themes to be improved, which some participants did not like.

Furthermore, the tone and length of the feedback were also perceived as sub-optimal, which is attributed to the fact that participants did not like the very verbose language of chatGPT. Participants noticed the overly enthusiastic and adjective heavy format chatGPT applies in their feedback more and this was perceived as unnecessary. One specific highlight in the iteration was that the compliment sandwich was utilized again. This format led to people that scored all points to receive a negative point, which more often than not was incorrect or redundant.

Lastly the feedback was seen as motivating (6 codes against 1 code). Especially high scores now received a highlighted example they answered correctly, which was found to be motivating and led to feeling more personalised.

4.1.5 Parametric Feedback Comparison and Evaluation

In order to evaluate the impact of generated feedback compared to the existing parametric feedback in the game, participants were asked to express their preference

between the newly generated feedback and the in-game parametric feedback. The results of this comparison are summarized in Table 4.2, which indicates whether participants preferred the generated feedback, had mixed feelings (denoted as "Mixed"), or found the generated feedback worse than the parametric feedback. In 5 instances, comparisons were not made because 3 participants did not complete this part of the evaluation.

Overall, the data reveals a strong preference for the generated feedback, despite occasional inaccuracies or "hallucinations" in the feedback, which occurred in 5 instances for feedback that was rated better. This preference suggests that the feedback generation prototypes positively influenced the gaming experience. However, the data also shows a variation in the reception of different prototypes. For instance, in the first iteration, 63% of responses indicated a preference for the generated feedback when removing the "No comparison" responses from the total cases, highlighting its initial success. In contrast, the second iteration saw a decrease in preference, with only 48% favoring the generated feedback under the same conditions. The third prototype demonstrated a significant improvement, with 89% of participants rating it as better, indicating a progressive enhancement in the feedback generation approach.

TABLE 4.2: Comparison Generated vs. Parametric Feedback and amount of Hallucinations in feedback

Comparison Generated vs. Parametric Feedback		
	Prototype 1	Hallucinations in Feedback
Better	17	2
Mixed	2	0
Worse	8	0
No comparison	3	-
	Total: 30	
	Prototype 2	Hallucinations in Feedback
Better	10	2
Mixed	6	2
Worse	5	1
No comparison	5	-
	Total: 26	
	Prototype 3	Hallucinations in Feedback
Better	23	1
Mixed	1	1
Worse	2	2
No comparison	0	-
	Total: 26	

Furthermore, the parametric feedback stayed consistent over the duration of the study. Because this feedback stayed consistent codes related to this feedback also stayed consistent over all iterations. Results of the analysis on this feedback are therefore presented together.

Out of 67 identified codes, participants most frequently commented on the feedback's length and its consistent nature (18 codes). The predominant sentiment was that the feedback was too short but more to the point. Participants that did not like

the positive tone of the generated feedback preferred this consistency more. However, there were concerns regarding the actionability of the feedback; 11 codes indicated that the feedback was not particularly useful and 6 codes related to motivation clarified that the feedback often provided positive reaffirmation without guiding specific changes.

The feedback did not utilize examples (2 codes), only in one instance for a specific episode this was mentioned. This also led to participants feeling the feedback was less specific (10 codes) and less personalised (3 codes). Due to the shorter length the tonality (11 codes) was found to be more strict and systematic, which some participants liked and others disliked; even stating that the feedback felt less human than the generated feedback.

4.2 Quantitative Analysis

4.2.1 Word Count Analysis

The first iteration of the feedback generation process, represented as Dataset 1, had a mean word count of 72.48 with a standard deviation of 30.98. This is almost double the word-count of 40 it should limit itself to. The boxplot in figure 4.4 for Dataset 1 displayed a wide range, suggesting a high variability in feedback length. The median, shown by the line within the box, indicated that most feedback fell around the 72-word mark, but there were instances of significantly longer feedback as illustrated by the upper whisker extending up to 160 words.

Improvements in the design process for the second iteration resulted in a more concise feedback generation, as evident from the statistical figures: a mean word count of 61.55 and a standard deviation of 11.04. This is well within the word-count of 80 which it should stay in. The boxplot for Dataset 2 showed a much narrower spread compared to Dataset 1, indicating a more consistent feedback length among samples. The absence of extreme outliers in this dataset underscores the enhanced consistency.

The third iteration, Dataset 3, recorded a mean word count of 66.08 with a standard deviation of 10.69. The distribution, as depicted by the boxplot, was similar to that of Dataset 2, showcasing tight grouping around the mean with minimal variability. Notably, there were a couple of outliers, as represented by the points beyond the lower whisker. One is from a negative play-through and one from a positive play-through, from different episodes. These outliers are thus probably due to natural variances in the model.

4.2.2 Sentiment Analysis

The Sentiment Analysis Test assessed the accuracy of the feedback's tone relative to gameplay outcomes over the three iterations. Initially, the feedback system performed well, with a 74% accuracy rate in the first iteration. However, there was a noticeable decline in subsequent iterations, with accuracy dropping to 64% and then to 60%. This trend suggests that while the initial feedback system was relatively effective in matching tone to gameplay choices, its ability to maintain this accuracy diminished over the iterations.

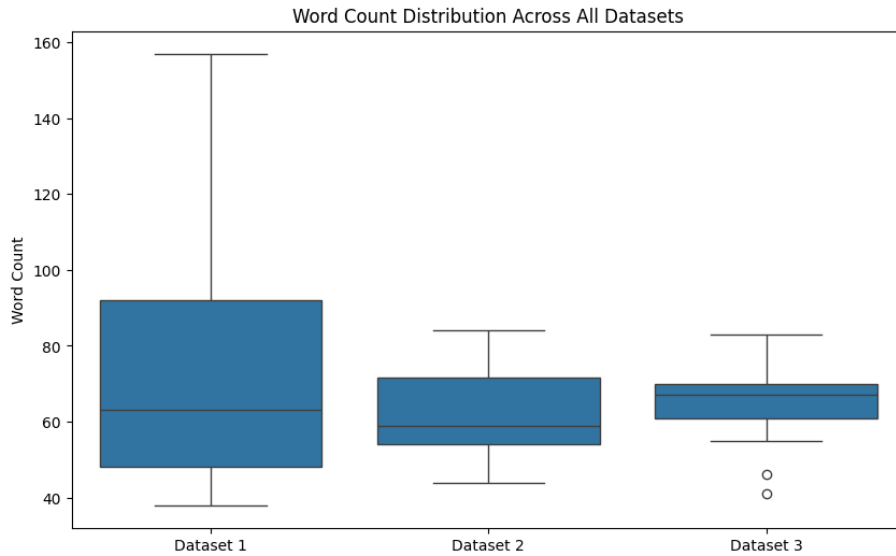


FIGURE 4.4: Boxplots of Word Analysis per Iteration

Title episode	Iteration 1		Iteration 2		Iteration 3	
	Episode played	Correct Sentiment	Episode played	Correct Sentiment	Episode played	Correct Sentiment
Meet the Team	5	4	4	3	5	4
Dropping the Ball	6	5	7	6	8	4
A House Divided	8	6	9	6	7	5
The Pressure Cooker	5	4	5	3	2	1
Case of Mondays	5	3	2	0	2	0
The Other Perspective	2	1	1	0	1	0
	74% correct		64% correct		60% correct	

TABLE 4.3: Sentiment Analysis of Feedback

4.2.3 Personalization Measure Analysis

This analysis explored how player choices impacted the feedback generated, using BERTScore. The regression analysis across the three iterations demonstrated that script similarity significantly influenced feedback similarity, as indicated by p-values less than 0.05 in all iterations. In Iteration 1, a significant positive slope (p-value: 0.012) indicated that as scripts became more similar, feedback also aligned more closely, suggesting a responsive and effective feedback system. However, Iteration 2 showed a significant negative slope (p-value: 0.019), meaning that increased script similarity led to less similar feedback, which could point to inconsistencies in how the feedback algorithm processed player inputs. In Iteration 3, the trend reversed back to a positive slope (p-value: 0.025), indicating improvements in the feedback system to again align closely with the scripts.

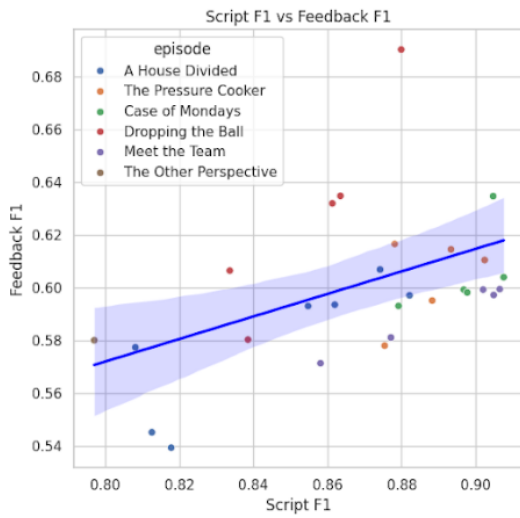


FIGURE 4.5: Regression plot of F1 score Feedback and Script prototype 1

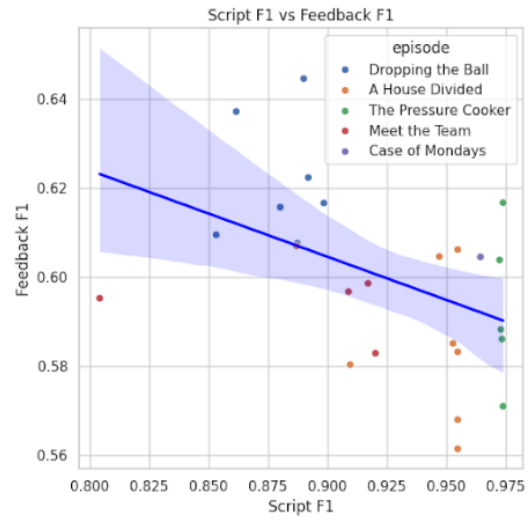


FIGURE 4.6: Regression plot of F1 score Feedback and Script prototype 2

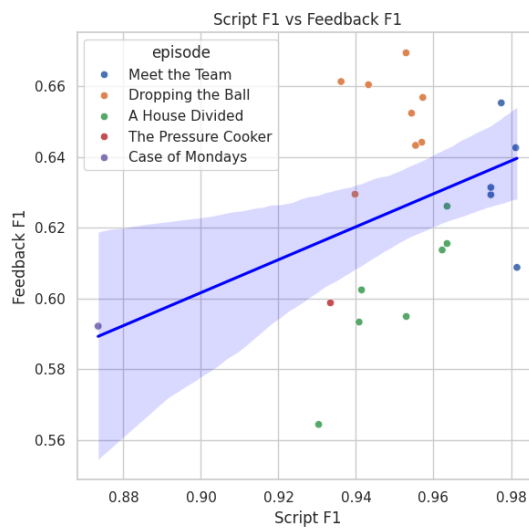


FIGURE 4.7: Regression plot of F1 score Feedback and Script prototype 3

Chapter 5

Discussion

5.1 Design Considerations

In this section, results are compiled and abstracted to explain the design decisions made during the development of the new prototypes and with it inform new practices. The initial design, detailed in the Methods Chapter, was constructed prior to results gathering, drawing on findings from the literature. The subsequent design considerations are informed by all the results presented in Chapter 4.

5.1.1 Prototype 1: Black-Box Approach

Considerations

Feedback from the first iteration showed promise and was shown to already produce actionable and personalised feedback in 64% of the cases, which was an improvement over the status quo. This black-box approach, did however face one major issue and that was its inconsistency.

As illustrated in figure 4.4, and reflected in the analysis of the codes, the length varied greatly. Despite this variability, participants generally appreciated the length and the depth it provided. The main inconsistency highlighted by participants was that not every generated feedback featured an example. Participants had a very positive association with these examples as across multiple codes examples were stated to have a positive impact. The examples however, needed to be of better quality.

Participants also noted that the tone of voice was too positive and suggested it could be more strict. As seen in table 3.1, 74% of the feedback accurately reflected the appropriate sentiment, whether positive or negative, based on the feedback score. This highlights that feedback sentiment should be improved.

5.1.2 Prototype 2: Naive Approach

Design

Following the initial design considerations, the second prototype, illustrated in Figure A.2, introduced a more complex system using prompt-chaining to achieve specific and consistent objectives. To enhance control over prompts and easily gather specific results, GPT API was prompted to give the responses in-between; "<tags>", which can easily be extracted using string parsing.

Furthermore, this iteration employed four prompts. The first generated initial feedback. This prompt closely resembled the baseline structure as seen in figure A.1, with the exception that the prompt is now more geared towards the goal of the episode instead of just the topic of the episode, which was hypothesised to make the feedback of better quality.

The second prompt elaborates on this feedback to make sure each feedback would have a specific example that fits that theory. Based on initial feedback generated, ChatGPT then selects the best example to go together with that feedback. So it chooses the example that fits the script the best.

The third adjusted the tone, as the tone was seen to not be fitting. This prompt tried to make the tone of the feedback more fitting and revised the feedback based on the example. The prompt was then fed into the last prompt. This refined the feedback for clarity and conciseness. It was also kept concise by requesting that the feedback stays within a specific word limit of 80 words, as the feedback length in the last iteration was generally found to be preferable.

Additionally, changes were made to the script incorporating all possible player choices to enhance the examples, by giving more context for explainability and make the examples even better by also being able to state what the answer should have been. To further enhance this additional information was provided to the system by stating which answers actually had an impact on the scoring and which did not. All other inputs were kept the same and were focused on capturing both the player experience goal and characterizing goal.

Considerations

Results from the second prototype show that the changes in the design did not have the desired effect and led to more incoherent and confusing feedback. Even though the feedback contained examples, these were not implemented well and the system did not have a correct idea of what examples to pick. This led to wrong or non impactful examples being chosen. These examples also made the feedback for higher scores to negative focusing on the one thing they did wrong or even hallucinating that one of the answers the player gave was a wrong one.

The results in figure 4.6 further showcase that the player choices did not have an impact on the feedback, highlighting that the example choice prompt probably did not perform well, or that even the whole structure of the chaining was incorrect. In the sentiment analysis it can be seen that the tone prompt did not have the desired effect and led to a decrease in correct sentiment.

Feedback length however became more consistent in this iteration, as seen in figure 4.4. The word count had a smaller standard deviation and 'whiskers', validating that prompt four worked.

Deeper qualitative evaluation highlighted that the feedback lacked reasoning and that it did not feel like a coherent story. This highlighted that the feedback needed a more empathetic or emotionally-driven spin to it, instead of only a summation of theory.

Furthermore, feedback was prone to over-fitting theory data to cases where it was not applicable, presumably to fill in gaps of choosing the wrong example. This shows

that the feedback was incohesive and not binding it with the episode participants just played.

5.1.3 Prototype 3: Heuristic/Reasoning Approach

Design

Despite enhancements, the second iteration exhibited issues with output coherence, prompting the development of a third iteration that can be seen in A.3. This iteration continued to employ prompt-chaining but emphasized a clearer reasoning behind feedback generation making the player experience goals and characterizing goal clear to the system.

This required additional narrative data, including detailed character traits of NPC characters, to improve role-playing elements and emotional state assessments. To address inconsistencies caused by overly broad theoretical inputs, a narrower scope was adopted, focusing on providing only the specific theory the player needs to learn for an episode. Changes were also made to the scoring/consequence function that had to explain if the question had impact on the score. This was done with numbers but now this would only be done with words.

Furthermore, the structure is completely overhauled, except for the last consistency prompt that was proven to work. The first prompt now assessed if the characterizing goal that the player had to learn. This reflects the base system but now only focusing on assessing. The second prompt focused on the emotional impact the player had on the character, and therefore assessed if the player experience goal was met.

The system then had to choose an example to base the feedback on. To further enhance this, the system chose a positive example, when the narrative goal was met, highlighting a correct choice the player made during the interaction. On the other hand the system chose a negative example when the player did not achieve the narrative goal, so when the NPC character felt a negative emotion after the episode. This insured that players with a high score got the positive reaffirmation they deserved and players with a low score got the help and explanation they needed to improve.

To make sure the system did not choose an example at random the scratchpad methodology was applied in the prompt. So, the system could first reason a bit about multiple answers before choosing one final answer and put that between tags. Furthermore, multi-shot prompting was used to make sure the examples chosen also featured the other answers.

From all these components the feedback is constructed focusing on giving the correct feedback for both positive and negative playthroughs and making sure the feedback is coherent with the correct theory and based in the narrative.

Considerations

Results from this design are a big improvement on the earlier prototypes, highlighting very high agreeableness with participants as can be seen in table 4.2, even better length consistency as seen in 4.4 and actual impact of the choices on the feedback as seen in figure 4.7.

Further qualitative analysis also highlights the actionability, specificity and personalization of the system. However, participants still had problem with the style and

tone of the feedback, which was indeed removed as a prompt, because it did not achieve its goal in design 2. Sentiment analysis also highlights that the tone of the feedback did not match the the score of the player in 40% of the cases.

Examples were also found to be lacking, specifically for higher scores participants wanted to have the exact answer they missed. However other participants liked that the feedback was more broad and tried to explain what themes you should work on or only gave personalised positive reaffirmation.

Giving away the right answer as two participants stated is a deeper design choice that is often discussed in *Serious Games Design*. Choosing to do this is up to the game-designer and might be very context dependent, even though participants would really like it it may not promote the actual learning process.

5.2 Research Findings

This research set out to gather insight into using generative AI for automatic feedback generation in serious games, to try to reduce design time, support increased complexity and better tailor responses to players. For this an iterative process was chosen to design and develop a system and directly learn from implementing and evaluating results. In the third prototype no new problems arose that were not found in earlier iterations indicating that the problems were not part of the design of the system, and it was chosen to make this prototype the final iteration. Insights from all results are discussed in this Chapter.

In comparison to the existing in-game feedback, the automatic feedback generation system was preferred in both the first and third prototypes, highlighting the success of these iterations. The key difference between the traditional parametric feedback and the generated feedback was its actionability. The automatic feedback provided more actionable goals, demonstrating its potential to guide players more effectively.

During the evaluation of the final prototype, it became clear that the generated feedback was more effective because it focused specifically on what went wrong, pinpointing the exact topic or theory that was not applied correctly. While some participants preferred the concise language of the parametric feedback, the final evaluation revealed that it was often too general and sometimes failed to address the actual issues. Addressing this with the parametric feedback would require incorporating more variables and possibly providing feedback for every possible combination of right and wrong answers—a challenge that highlights the scalability advantage of automatic feedback generation.

Reviewing the reflection of the designs, what becomes apparent is that the problems around the tone and style of the feedback stay consistent across all three iterations. While it was tried to overrule and influence the tone directly in the prompt and even with a separate prompt, the positive tone from the system prompt baked into chatGPT is hard to overrule and causing it to provide too positive feedback to participants.

Furthermore, the highly verbose language use with a lot of adjectives, gave participants a negative experience. Highlighting this made the feedback less concise and made it feel like AI. In addition to this, the compliment sandwich style was never specified and ChatGPT automatically uses this for feedback, which is not always preferable and gives the system less flexibility.

These findings indicate that ChatGPT, while having an easy to implement API, might not be the best model to generate feedback in serious games as it does not have a lot of customization in terms of tone. Prompting has shown to not have a big impact on this. Other less restricted models like the open-source LLama, might perform better in this regard as this model has less restrictions and is more customizable. However, fine-tuning of GPT models might also solve this problem.

Feedback generated by chatGPT is however actionable, and applicable in many cases. As seen in design one, mapping the information needed to explain the characterizing and player experience goal of the game and feeding that into the system gave actionable feedback about the topic, fitting the narrative. This can then easily be enhanced with one other prompt to make the feedback more concise in length. A simple system only focusing on providing actionable feedback, could easily be made in this manner.

However if the serious game needs more personalization and more specific responses more reasoning is required. As seen in the Results chapter this requires providing the system with the tools to reason and build the feedback from the ground up. The system needs to both understand and evaluate the characterizing goal and the player experience goal to effectively generate feedback. Examples can then be better chosen to explain why these goals were met or not. GPT does not have this reasoning inherently, as seen in design 2 where a specific example was tried to be implemented with only the feedback as guidance, and adding the elements highlighted in evaluation 1. However, this resulted in worse not coherent feedback that was out of scope.

As seen in literature, designing feedback consists of three elements, feed-up; what are the goals, feedback; what progress is being made to this goal and feed-forward; where to next (Hattie and Timperley, 2007). The proposed prototype in design A.3, effectively captures these three elements as seen in the generated results, which might explain the high approval rate in table 4.2. This further supports that this goal-oriented approach works well, highlighting the need for a system to understand fundamental elements of feed-up and feedback.

Furthermore, when designing an automatic feedback system for serious games, it's crucial to map inputs effectively and ensure all information is interpretable by a Large Language Model (LLM). During the development of the three prototypes, various information sources were gathered and utilized as inputs for the feedback system. By grouping information based on characterization and player experience goals, the system becomes more flexible and easier to understand for both designers and the system itself. In Prototype 2, an attempt was made to enhance reasoning by incorporating numerical data. However, this approach negatively impacted the LLM's reasoning process, as LLMs struggle to interpret numerical data effectively. Instead, providing clear explanations with specific contextual information proved to be more successful. Nevertheless, extensive testing and reformulation are necessary to ensure the system correctly interprets the inputs.

Additionally, scoping the information is critical. As observed in Prototype 2, when utilizing multiple prompts for more advanced systems, excessive information can disrupt the feedback system, leading to incorrect feedback. Ensuring that the right information is provided at the appropriate moment is essential. In this system, all inputs were hard-coded. Scoping the information to only the specific theory that needed to be applied to achieve the characterizing goal seemed to work the best, as illustrated by prototype 3. In contrast, the player experience goal information was

given all context information off previous episodes, which did not lead to problems. Which highlights that specificity and context for both goals might be differently defined.

The aim of this research to support increased complexity also meant that the design had to be more flexible and generalisable for multiple games. While not directly tested in this research, design choices were made with this kept in mind. If a system for this specific game or similar closed-question narrative games were to be developed, more reasoning can be added by dividing the system further based on what the numeric outcome of the level was. Creating a Neuro-Symbolic AI system that can specifically tailor feedback per score, in this game star-rating, and meet the specific criteria of each rating. Combining the power of human reasoning and existing knowledge of the already developed game, with the power of flexible and in depth AI generated feedback, tailoring the game to the player Hitzler and Sarker, 2022.

Lastly, results from the last iteration as showcase that generated feedback almost reached a 89% approval rate over static parametric feedback. In the three cases that were labeled as worse, hallucination was detected. All 4 hallucination cases found also came from the same episodes. This highlights that there might be a deeper problem. One episode had to choose from context interaction type, which was represented differently in the script than other interaction types. Closer inspection on the feedback generated for these types highlights that this format is not something that ChatGPT could understand as it was missing important context from the answer that was chosen. Another problem this episode has is that Olivia, the character you interact with, is always happy, even if you give her wrong advice.

Similarly, the episode after that got unfitting feedback. Upon revisiting this episode it became clear that the tone of this episode is always negative. Even when you help Erika, the tone is still hostile and therefore the player experience goal of helping your teammates solve their problems and making them feel differently is not that clearly achieved.

Future serious games that want to incorporate automatic feedback generation using generative AI, need to pay attention to generative AI from the start of game creation. Enough information for both the player experience goal and the characterizing goal needs to be made clear, focusing on clearly defining the goal of each level, and specific information on how to reach that goal. AI needs to have enough context information so it can reason with it itself. It needs to be in a clear format with a lot of context, as it otherwise does not understand it and will fill in the gaps itself resulting in hallucinations.

5.3 Limitations

While this study provides valuable insights into the use of generative AI for automatic feedback generation, several limitations should be considered when interpreting the findings.

Firstly, the dual-platform development introduced limitations in the overall research design and development. Optimizations for Android and iOS hindered the feasibility of a simpler web build that would have made sharing the game easier. Additionally, stricter privacy regulations on iOS required approvals for new releases and features, presenting unnecessary hurdles. Therefore, the development of the prototype was limited to an Android build only.

Because no easily shareable iOS build could be made, iPhone users who participated in the online experiment played the game via remote access on their PCs, which is a very different experience from playing directly on a phone. Some participants misclicked or experienced issues with lag and stream quality.

Additionally, because the game was streamed directly from the Unity editor, in two cases, test data used during the development of the game got mixed up in the scenario script of the participants, resulting in a hallucination for one participant in the second iteration. However, since this participant still provided relevant feedback, this data entry was not removed.

Moreover, remote access and recording were only possible through Zoom, and with the basic free plan, the research sessions could only last 40 minutes. After that, the participant was forcibly removed from the call by Zoom, which resulted in two participants being unable to finish the comparison. Another participant did not have enough time to complete the comparison due to other commitments and had to leave early.

Two participants in the second iteration experienced game bugs during gameplay, where some of their answers were not correctly stored in the script due to changes made in the design. This resulted in error messages on the screen and, even worse, led to them receiving inaccurate feedback, including some hallucinations. Revisions were made between research sessions to address these issues.

Furthermore, while sentiment analysis tests are informative on paper, they are not the most reliable in this case, as qualitative data also highlights that motivational or positive text can be effective even when receiving a negative score. Since an arbitrary threshold for positive or negative sentiment was chosen for this test, it may not have been subtle enough to accurately reflect whether the score conveyed the correct sentiment.

Participants who were directly playing for a negative score also had a different experience compared to those who played normally for a positive score. The feeling when receiving feedback was harder to assess, as most participants were relieved when they received a low score. However, since this research was predominantly carried out with experts directly from the field, they were able to dissociate this feeling from their answers.

A more significant limitation was that, in each iteration, multiple participants reported that they did not remember what they had answered. Even though the episodes were very short, participants highlighted that they forgot which answers they gave, making it harder for them to evaluate the examples.

Furthermore, participants encountered one prominent issue with the game: multiple participants stated that none of the choices they could make fit their play style, leading to a less personalized and overall worse experience. In some cases, participants also chose specific answers expecting them to have a different meaning or connotation.

5.4 Future Work

In this section, there are two types of future work described: future work directly gathered as results from the expert evaluation and follow-up work based on the

main findings of this research. The exploration of qualitative data has garnered insightful suggestions from participants. These ideas, collected over three iterative sessions, provide a roadmap for further enhancing automatic feedback generation and the use of generative AI in serious games.

5.4.1 Recommendations from Participants

Participants in the first iteration emphasized the need for nuanced and accessible feedback mechanisms. Suggestions included providing links or buttons that point to further detailed explanations, and ensuring feedback is stored within the app for future reference. The concept of 'integral feedback' emerged, advocating that feedback should be reinforced into later feedback moments so the system can acknowledge player growth.

Furthermore, the notion of further personalizing feedback was strongly favored. This could involve formatting feedback in various styles or having it delivered by in-game characters, which could make the feedback feel more tailored and engaging. Additionally, enhancing the input of the system AI with more dynamic numerical metrics, like star ratings, and other performance measures such as the speed of gameplay or the duration of inactivity, could provide users with a clearer assessment of their performance and further differentiate feedback based on individual behaviour.

In line with future work of Gallotta et al., 2024, to help inform further personalization, more research needs to be conducted around player modelling, focusing on modelling, what kind of feedback participants like, and how to measure that. A study highlighting feedback styles with specific metrics to measure what participants like could then be conducted.

Building on the feedback from the first iteration, the second iteration introduced ideas focused on interaction and responsiveness. For example, with generative AI, there could be an interactive dialogue about feedback specific details in the feedback, or an open dialogue to dispute received feedback. Participants wanted to better understand why certain feedback was given and liked the idea of utilizing the chatbot nature of GPT to enhance this learning effect.

It also became apparent that including a 'thinking' animation when the AI is generating a response is a must, especially when the generation takes some time. This is presumed to make the system appear more responsive and less confusing. The introduction of loading screens and a mechanism to indicate that feedback is being generated was also suggested to enhance personalization. Also explicitly mentioning that the feedback is being personally tailored would increase the feeling of personalization.

In the final iteration, the discussion on future work highlighted the importance of addressing accessibility and inclusivity, such as offering shorter or differently formatted feedback for players with reading impairments. The use of open-ended game scenarios was also suggested, where instead of choosing between answers players could write their answer in free-form text. In addition to this feedback, metrics tailored to match the user's input length, verbose language use, etc. could be extracted to further enhance the feedback.

Finally, incorporating elements like streaks or other gamification tools in the feedback was suggested to maintain player engagement. The personalization of feedback, clearly stating that it is tailored for the player, and providing engaging activities during waiting periods, like tips or trivia, were additional points to enhance the overall user experience.

5.4.2 Future Research

In addition to the points provided by expert evaluation, there is future work directly following this research. First and foremost, the design needs to be tested with multiple AI models, like Claude, LLama, Gemini, etc. to evaluate if the heuristic approach works across multiple models or is specific to GPT. In addition to this the models performance and feedback need to be assessed to see which performs better in terms of tone, as participants did not like the stock, helpful tone-of-voice of GPT.

Although the GPT API offers good integration and is a leading platform for novel and innovative LLM tools, future research needs to focus on ways to circumvent the helpful tone-of-voice of GPT, either through fine-tuning or prompting, to better prepare this model for a wider variety of applications.

To further test the generalizability of the design, it needs to be tested on other serious games in different knowledge domains. GPT knows a vast amount of information, but it performs differently on different topics (Achiam et al., 2023). The context of the game 'Take 5', is soft-skill training, which is a very general domain. Feedback generation might struggle with more specialized domains for which GPT was not trained on with extensive data.

Looking further into the future of game design, free-form text, as highlighted by multiple participants, might become the new standard for educational games. Not only would it solve current problems some participants faced during gameplay, such as not finding any answer that fits their management style, but it could also enhance player modeling and subsequent feedback personalization. The proposed design should be tested with this as well. However, meticulous care is needed, as participants will not have a star rating or existing feedback to compare their generated feedback to.

Lastly, this research was primarily interested in the applications of this technology in serious games, not the ramifications and impact it has on the educational sector or beyond. However, literature has highlighted multiple concerns that future research needs to address before AI can be comfortably applied in all systems.

Deploying GPT raises ethical and privacy concerns. Managing sensitive student data and ethically using these technologies is crucial, necessitating stringent measures against biases in language processing and ensuring data privacy (Yan et al., 2023; Kasneci et al., 2023a; Owan et al., 2023; Dwivedi et al., 2023; Pechenkina, 2023; Mørch and Andersen, 2023). Additionally, addressing potential biases in language processing and AI algorithms is vital to ensure fairness in educational outcomes (Yan et al., 2023; Baskara, 2023; Owan et al., 2023).

Furthermore, there remain concerns regarding the reliability and replicability of ChatGPT's outcomes in diverse educational settings (Yan et al., 2023; Rasul et al., 2023). Ensuring consistent benefits of ChatGPT across different contexts is crucial for

its widespread adoption. Finally, effectively integrating ChatGPT in educational settings requires careful design and implementation, considering ethical usage, aligning with learning theories, and preserving context in educational content (Baskara, 2023; Mørch and Andersen, 2023).

In addition to this sustainability, the environmental impact of training and deploying LLMs is also underrepresented in research (Gallotta et al., 2024). For closed-ended questions games, similar to 'Take 5', an approach to reduce this impact could involve pre-generating feedback options rather than generating them on the fly. This strategy would involve developing a comprehensive set of potential feedback responses that can be reviewed and refined before implementation in the game. Such an approach could decrease waiting times for feedback, ensure higher quality and consistency of responses, and reduce computational demands during gameplay, although it would require considerable upfront development work. Further research into whether this kind of approach would impact sustainability is preferable.

Chapter 6

Conclusion

This thesis set out to explore the potential of generative AI in creating automatic feedback systems for serious games, aiming to improve feedback quality, reduce design time, and handle increased complexity. The research was guided by three primary questions: What constitutes effective process-based feedback for automatic feedback generation in serious games? What are the key design considerations for integrating generative AI into serious games for automatic feedback generation? How can generative AI be utilized to develop an automatic feedback generation system within serious games?

To address these questions, an exploratory sequential mixed methods design grounded in Wieringa's design science framework (Wieringa, 2014) was employed. The methodology involved iterative prototype development and evaluation across three main cycles, with each iteration building upon the findings of the previous one to refine the design and improve the system's performance.

A mixed deductive and inductive thematic analysis was performed across the three iterations to identify key elements of effective feedback. This informed the answer to the first research question. Literature revealed that the content of effective feedback in serious games should be actionable, specific, and personalized, aligning with both the characterizing and player experience goals of the game. Inductive coding in addition to this found that feedback in serious games also has a motivational effect and highlights this as an extra requirement. Overall, this analysis concluded that Generative AI, such as GPT-4, can generate actionable feedback when provided with detailed context and narrative goals, though customization for tone and specificity remains challenging.

As to the second research question, design considerations were found and implemented across the iterations. The first iteration, referred to as the black-box approach, focused on mapping input and categorizing essential aspects needed for feedback generation. While the initial prototype provided actionable and personalized feedback and performed better than the old feedback already did in 64% of cases, it faced issues with inconsistency and tone. The second iteration, the naive approach, introduced prompt-chaining to achieve specific and consistent objectives using four prompts to enhance control over feedback generation. Despite improvements, this iteration still lacked coherence and reasoning, and tone issues persisted. The third iteration, called the heuristic approach, emphasized clearer reasoning and narrative data to improve coherence and emotional alignment. This prototype integrated additional narrative data and focused on the specific theory needed for each episode, significantly improving feedback quality and participant satisfaction.

Considerations are further built on-top of quantitative analyses, to validate the generated feedback. The first test, a word analysis, evaluated the consistency of feedback length, revealing that prompt chaining was an effective solution to this problem. The second test, a sentiment analysis, assessed the tone of the feedback in relation to player performance, showing that across all three iterations sentiment did not achieve correct results, where both positive and negative feedback had the wrong sentiment. The third test, a semantic regression analysis was performed on two corresponding BERTScores, measuring the similarity between player actions and on the other hand the similarity between the generated feedback generated feedback, highlighting areas where the AI effectively captured player intent, in the feedback and areas where it fell short. This showed that regarding the design considerations both the black box the heuristic approach had captured the player behaviour well in the feedback.

Regarding the third research question, the integration of generative AI into serious games requires clear goal definitions and sufficient contextual information to avoid AI hallucinations and ensure relevant feedback. A multi-prompt approach can help manage feedback quality, but tone adjustment require careful design and may benefit from fine-tuning or using less restrictive models than GPT. This research demonstrated that generative AI could be effectively utilized to develop feedback systems by applying a design that captures the player experience and characterizing goal and gives AI the means to reason with these goals. Designing the correct feed-up and making sure tailored feedback in the form of examples is provided, AI is able to give correct feedback and provide actionable feed-forward. With proper context and narrative integration as input, AI-generated feedback could achieve high approval rates.

In terms of future work, several areas were identified for further exploration. These include testing the design with multiple AI models like Claude, LLama, and Gemini to evaluate the heuristic approach's effectiveness across different models, and exploring fine-tuning techniques to better control the tone and style of feedback. Additionally, this research suggests testing the design on other serious games with different knowledge domains to assess its generalizability. The use of free-form text input and the integration of adaptive and personalized feedback systems are also recommended as promising directions for future research.

In conclusion, this research contributes to the fields of serious games and automatic feedback generation by providing a structured approach to integrating generative AI, highlighting both its potential and areas for improvement. The insights gained can guide future developments in creating more effective and efficient serious games, ultimately enhancing the learning experience through better feedback mechanisms.

Appendix A

Prototype Designs

In this appendix the designs of the three iterations can be found. In addition the specific prompts that were used in each task/ sub-tasks are provided. Prompts are structured in as `c#` strings, and normal string operations to optimize the usability and readability of these are utilized. Further inputs are explained in the methods or discussion chapter.

A.1 Prototype 1: Black-Box Approach

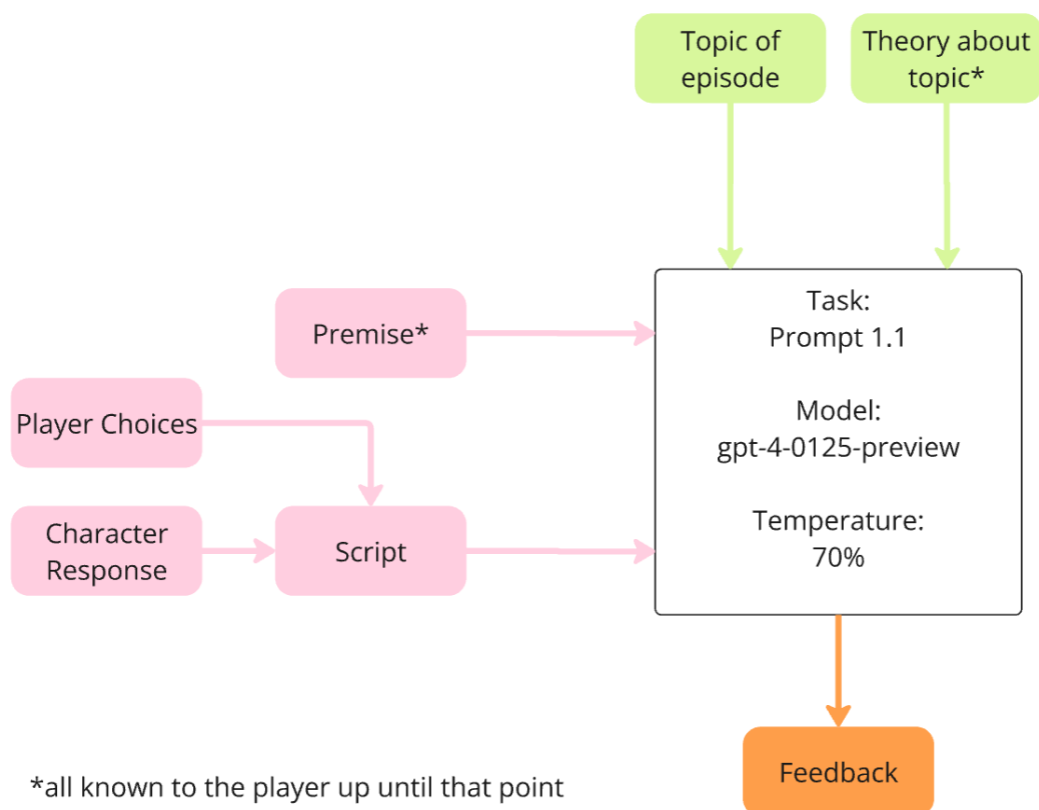


FIGURE A.1: Design Prototype 1: Black-Box Approach

A.1.1 Prompt:**A.1.2 System prompt:**

You are a coach, training people in leadership skills, in this episode, the player needs to learn `SaveDialoguePrompt.GetCurrentTopic()`. Give feedback on `SaveDialoguePrompt.GetCurrentTopic()`, based on the given context of the questions/answers the player asked/gave, directed at the player. limit the response to words words. Be specific, actionable and make the feedback personalised, you can be strict. The player has limited time and is unable to ask more questions, don't give feedback to delve deeper or ask more question! Your personality: Talkative, Enthusiastic and Care-free";

User prompt:

\$"The premise of the story so far is: premiseTextThis is the JP information you already provided before the interaction: jpText";

A.1.3 Generated Feedback:**Participant 1.1**

Meet the team: Your approach to asking open-ended questions, like with Olivia and Paul, shows promise in engaging others. However, try to reflect on or paraphrase their responses to demonstrate deep listening. This will make your interactions feel more thoughtful and less transactional. Great start, though! Keep it up!

You initiated the conversation with empathy, which is fantastic! Remember to use the SEE framework next time: specify the situation, explain its effect, and express the desired behavior. This approach will make your feedback even more actionable. Keep up the good work!

Fantastic job navigating that conversation with Erika! Your open-ended questions gently guided her to open up about her struggles, showcasing your powerful questioning technique effectively. Plus, your patience and silence allowed her space to reflect and share more deeply, embodying deep listening. Remember, affirmations like That sounds pretty bad! not only validate feelings but also encourage further sharing. Great balance of empathy and action-oriented support at the end! Keep it up!

Participant 1.2

You're off to a vibrant start, diving right into getting to know your team! Your questions to Olivia and Paul were particularly insightful, showing genuine interest in their experiences and ideas. However, remember, deep listening involves reflecting on what's said to build a deeper connection. Try echoing sentiments or asking follow-up questions based on their responses next time. It's not just about the question but how you engage with the answer. Keep up the enthusiasm, just channel it into more engaging responses too!

You initiated the conversation with Erika positively and made a good effort to structure your feedback around specific issues. Remember, specificity helps! For future interactions, dive deeper into the SEE framework by explicitly stating the effects of

Erika's actions to paint a clearer picture before moving to solutions. Keep up the enthusiasm!

Hey there, superstar! You dove right into the deep end with Erika, didn't you? Your open-ended questions were like keys unlocking doors – brilliant strategy! “Where would you say that's coming from?” – Chef's kiss on that one! You gave Erika the space to think and express herself, showing off some top-notch deep listening. Just a tiny nudge, though: While empathy was shining through, ensuring a bit more focus on actionable steps earlier could steer the convo towards solutions a bit snappier. But hey, you're juggling those skills like a pro! Keep it up!

Participant 1.3

Wow, you've got a knack for engaging! Your questions were open-ended, encouraging others to share more about themselves. However, remember to dig a bit deeper next time to truly leverage powerful questioning and deep listening. Great job connecting on personal interests, but aim for questions that also reveal their professional insights or challenges. Keep it up!

You started strong by inviting Erika to reflect on her performance, which sets a collaborative tone. However, remember the SEE framework; you missed elaborating on the specific effects of her actions and concrete examples. Next time, ensure you express the desired behavior more clearly to guide Erika towards improvement. Great attempt at keeping the conversation open-ended for future discussions, though! Keep it up!

You nailed the art of Powerful Questioning by asking open-ended questions that encouraged Erika to share more about her situation. Your Deep Listening skills shone through as you gave Erika space to think and express herself, showing empathy and understanding. Remember, acknowledging her feelings directly can further validate her experience. Great job in navigating the conversation! Keep up the good work!

Participant 1.4

Your approach to questioning was a good start, showing a mix of personal and professional interest. However, try to avoid making comparisons about differences too quickly; it might create unintended barriers. Celebrate the diversity in responses more openly to foster a welcoming atmosphere. Keep up the enthusiasm!

Your approach shows promise, especially in trying to engage Erika and planning follow-ups. However, you missed opportunities to clearly articulate the 'Effect' and 'Express' desired behaviors in a supportive manner. Next time, focus on empathy and specific examples to foster a more productive dialogue. Aim for clarity and understanding, while ensuring Erika feels supported throughout the conversation. Keep it up, but let's add more precision and warmth next time!

Participant 1.5

Your approach to opening the conversation was solid; acknowledging Erika's time constraints but emphasizing the importance of the discussion was a good balance. However, your questioning leaned a bit towards assumptions, particularly with “You're having trouble at home, is that it?” It's crucial to maintain open-ended questions without implying an answer to foster a more open dialogue. For example,

What's been occupying your time lately? might have encouraged Erika to share without feeling cornered. In the listening front, you did well to allow Erika space to express her frustrations, indicating you were practicing deep listening. Yet, the transition to suggesting she focus on work felt a bit abrupt. Remember, deep listening also involves empathizing with the speaker's situation before rushing to solutions or advice. Short, keep fostering open dialogue with neutral, open-ended questions and ensure your responses fully reflect an understanding and consideration of what the other person has shared. Keep it up; you're on the right track!

Your approach with Erika showed a good blend of concern and practical suggestions, but there were missed opportunities for deeper questioning. Instead of immediately suggesting solutions, explore Erika's feelings and challenges more. Ask "How does this situation make you feel?" to connect on an emotional level. Your suggestion to delegate to Olivia was spot-on, yet pushing Erika to reflect on her reluctance could have been more insightful. Try questions like, "What concerns you most about delegating to Olivia?" to understand her hesitation. Your listening was attentive, but remember, deep listening involves exploring the unsaid as much as the said.

Your use of empathetic statements ("You look a little shaken up") is commendable, fostering a supportive environment. However, your questions could delve deeper into understanding Olivia's perspective on her relationship with Erika, beyond the immediate incident. This could reveal underlying issues contributing to their conflict. Remember, powerful questioning is about unlocking new insights, not just gathering surface-level information. Aim to explore feelings, needs, and possible solutions more thoroughly next time. Keep up the good work on building rapport and encouraging proactive steps!

Participant 1.6

Your questioning technique started strong, indicating a desire to understand Erika's situation beyond surface-level issues. However, pressing for personal details might have felt intrusive rather than supportive. In future, focus on empathetic listening, allowing Erika to share at her own pace. Aim to make her feel understood and supported, not interrogated.

Your approach was empathetic and you effectively used open questions to let Erika express herself. However, your initial attempt to address her enjoyment at work felt a bit direct. Next time, try framing observations without assuming feelings. Your suggestion about delegating to Olivia was practical, but ensure to follow up on Erika's apprehensions more reflectively. Good job on encouraging planning and delegation! Keep fostering these conversations to support Erika's growth.

You started strong by creating a comfortable atmosphere for Olivia, showing empathy with the coffee gesture. Good job using powerful questioning to delve deeper into her frustrations. However, your focus leaned heavily on the incident details rather than exploring the underlying issues between Olivia and Erika. Next time, aim to balance understanding the specifics with uncovering the root causes of their conflict. Encouraging Olivia to take the initiative was a positive step towards resolution. Remember, deep listening involves not just hearing words but also noticing the emotions and underlying messages. Keep it up, but try to probe a bit deeper next time!

Participant 1.7

Great energy in connecting with the team! However, your questions sometimes veered towards personal rather than exploring their work style or how they contribute to the team. In leadership, aim to understand their professional selves first. Keep the enthusiasm high but focus more on relevance to their roles!

You jumped in with strong labels (non-performing employee; unfocused) which can be confrontational. Remember the SEE framework: start with a specific situation, like the onboarding process, then express the effect, and suggest a desired behavior. Aim for a supportive tone next time!

Your approach started strong by setting a clear timeframe, which shows respect for Erika's time constraints. However, your questioning seemed a bit leading and assumptive, particularly with, "There must be something else that's bothering you, what is it?" and "Do you think you have your priorities right?" These questions might put Erika on the defensive. While you did well to encourage openness, a more effective strategy could have been to focus on how her situation affects her ability to meet commitments and explore support she might need, without implying judgment on her priorities. Try to maintain a balance between expressing concern and avoiding assumptions about what's best for her situation.

Participant 1.8

Your approach showed genuine concern and a good attempt at deep listening, but your questioning could be more exploratory to uncover the root causes of Erika's stress. Instead of suggesting solutions right away, focus on empowering Erika to identify her own solutions. Encourage self-reflection to foster her independence and confidence. Keep up the enthusiasm but remember to guide, not prescribe.

You started strong by offering Olivia comfort, showing empathy. However, focusing on specifics like the coffee spill might have diverted from deeper issues. Next time, aim for questions that explore underlying work relationship dynamics. Good job encouraging proactive communication! Keep it up, but dive deeper.

You started strong, inviting Erika to share her perspective, which aligns well with powerful questioning. However, your response "Don't you think you're being a little harsh?" could potentially shut down openness rather than encourage deep exploration. Instead, aim to maintain neutrality and curiosity, like when you asked about other situations where Olivia made her feel that way. This approach fosters a safe space for genuine reflection and insight. Also, remember to listen deeply to what's not being said and explore those silences with empathy and openness. Keep practicing to master these skills!

Participant 1.9

Your use of open-ended questions effectively drew out Erika's underlying issues, demonstrating excellent powerful questioning technique. However, your responses could show deeper empathy to enhance the connection. Remember, acknowledging emotions explicitly can make the speaker feel truly heard. Great job maintaining focus and keeping the conversation productive within the time limit. Keep it up!

Alrighty, let's break it down! Your initial approach to check in on Erika and asking about her rest period? Spot on! You started with empathy and genuine curiosity. However, when Erika mentioned her energy levels and work enjoyment, you could've dug deeper. Remember, powerful questioning is not just about the what, but also the why and how. Asking Why do you think your energy levels are lower? could've opened up more insights. suggestion to delegate and prioritize was fantastic! Encouraging her to think about who can support her and emphasizing long-term benefits showed your carefree yet strategic thinking., remember, deep listening isn't just about hearing words; it's about sensing the unsaid. Erika's hesitation about delegating to Olivia could've been explored more with, What specific concerns do you have about delegating to Olivia? That would demonstrate deeper empathy and understanding., you're doing great! Keep focusing on not just the content of what's

Your approach to powerful questioning really helped Olivia open up and reflect on her situation with Erika, nicely done! You successfully navigated the conversation towards self-realization for Olivia, which is a key leadership skill. However, let's also remember the importance of deep listening. While you did offer support and a listening ear, ensuring every response directly addresses Olivia's feelings and concerns can further strengthen your connection. Keep up the enthusiastic energy and continue to refine these skills!

Participant 1.10

Your approach to powerful questioning was effective in uncovering Erika's challenges and her thoughts on delegation and planning. You showed empathy and understanding, which is key. However, try to probe deeper on her reluctance to delegate to Olivia - what specific concerns does she have? Your deep listening allowed Erika to open up, but remember, summarizing her feelings and reflecting them back can deepen her reflection and your understanding. Great job on encouraging delegation and planning, but next time, ensure Erika feels more supported in addressing her concerns about Olivia's experience. Keep up the good work!

Your questioning guided Olivia to self-realization, which is fantastic! However, try to balance between exploring her feelings and encouraging actionable steps earlier. Deep listening was evident, but remember to also reflect back to ensure understanding. Keep it up!

Your powerful questioning nudged Erika to share specifics, which is great! However, your deep listening could improve by summarizing Erika's points before moving on, ensuring clarity and showing empathy. Keeping it concise and focused will boost understanding on both ends. Keep it up!

A.2 Prototype 2: Naive Approach

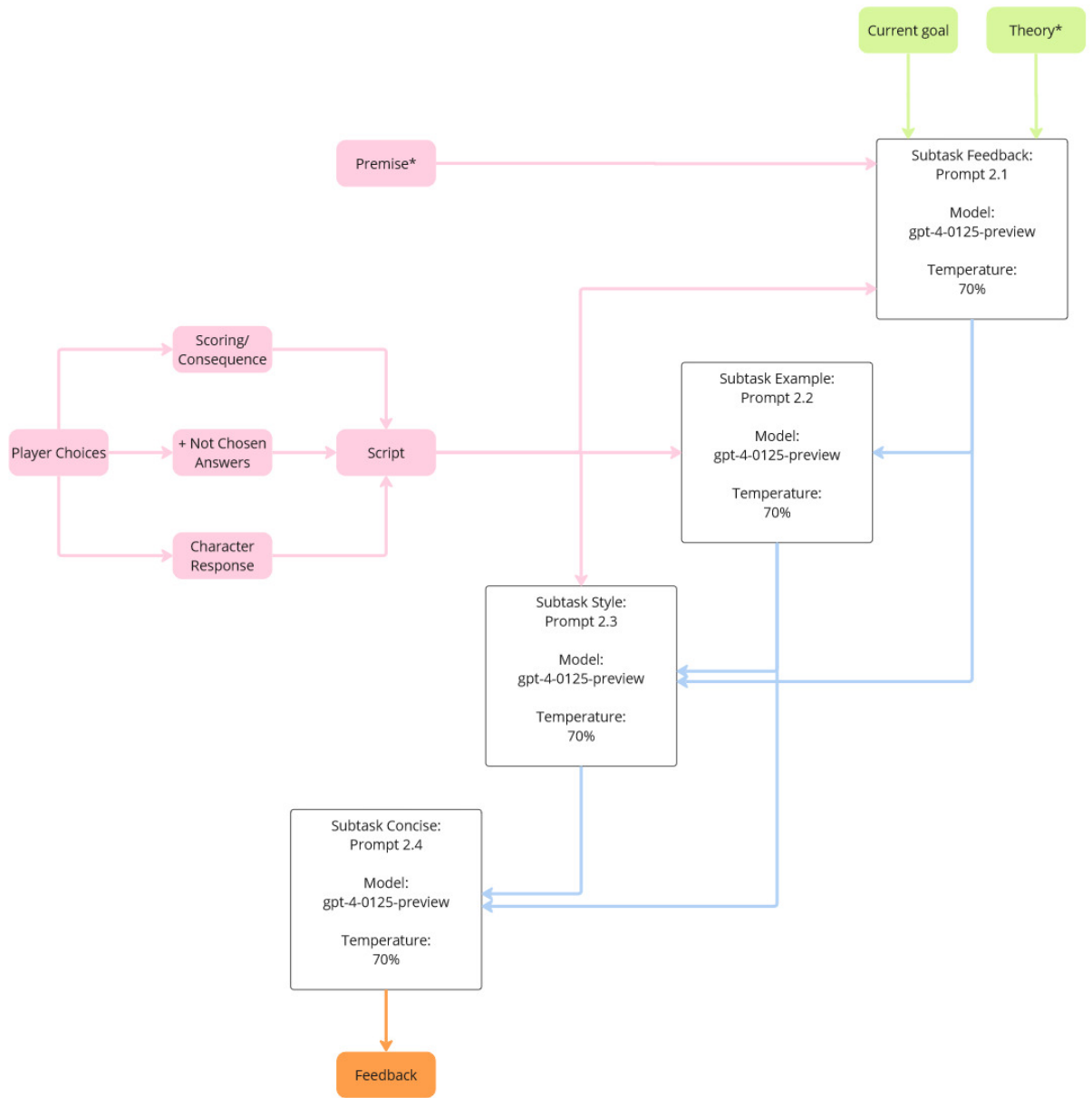


FIGURE A.2: Design Prototype 2: Naive Approach

A.2.1 Subtask 1: Feedback

A.2.2 System prompt:

```
string subtaskFeedback = @$"You are a leadership coach providing feedback to a
player who has just completed a training session on the following topics: <learn-
ing_goals> SaveDialoguePrompt.getCurrentGoal() </learning_goals>
```

In this training session, the player could only select a limited number of responses in each scenario.

Reflect on the choices the player made found in the <player_script> </player_script> tag

Encourage the player to review the relevant leadership theory and consider how they can apply it more effectively in future scenarios. Remind them that leadership is a skill that requires ongoing practice and reflection.

End your feedback on a positive note, acknowledging the player's efforts in completing the training session and their willingness to learn and improve their leadership skills. Emphasize that making mistakes is a natural part of the learning process and that the key is to learn from them and continue growing as a leader.

Make sure that when the player tries to play the scenario again they can apply the feedback you give them and perform even better.

Provide your feedback in a clear, concise manner, focusing on the most important points that will help the player improve their leadership skills. Write your feedback inside <feedback> tags.";

A.2.3 User prompt:

```
string input3 = @"<player_script>" + userInput + "</player_script> and here is the
leadership theory the player has learned so far: <theory>" + jpText + "</theory>
and the story so far <premise>" + premiseText + "</premise>";
```

A.2.4 Subtask 2: Example

A.2.5 System prompt:

@\$"Please carefully read through the player's script and the theory. Then complete the following tasks:

The player recieved feedback within the <feedback> </feedback> tag on their performance in the training episode. You need to identify in the <player_script> </player_script> the player's biggest mistake according to this feedback.

1. If the player got every single question wrong and scored a total of 0 points, state ""The player got every question wrong."" If the player got every single question right and scored a perfect 10 points, state ""The player got every question right."" Otherwise, do not comment on their overall score.
2. If the answers given by the player states that it had no consequences, this answer has no impact on the story or the points these answers are not considered for the biggest mistake answers

3. Identify the question where the player made their biggest mistake according to the theory. Explain in a `<biggest_mistake_reasoning>` tag why this was the player's biggest mistake based on the theory, also give add all other answers the player could have given in, so the answers the player did not choose to this tag and state which the player should have chosen.

Put your final chosen answer in the `<biggest_mistake_reasoning>` tag with your explanations.";

A.2.6 User prompt:

```
@"Here is the script of the player's answers and scores from the training session:
</player_script> " + userInput + " </player_script> And here is the leadership theory the player is has learned so far: <theory>" + jpText + "</theory> and everything the player has learned so far: </lessons>" + jpText + "</lessons>";
```

A.2.7 Subtask 3: Tone

A.2.8 System prompt:

@\$"magine you are a feedback coach in this serious game designed to teach leadership skills to players, this game is playfull but used in a professional corporate setting. Your role is to observe the player's actions and provide them with feedback to help them improve.

The feedback you need to reformulate to the player is within tags: `<feedback>`
`</feedback>`

As the coach, you have a talkative and enthusiastic personality. Your feedback should be delivered in an engaging, game-appropriate style that matches the player input.

Please reformulate the feedback in this style, while still ensuring the substance and meaning remains the same. The feedback should be:

- Actionable: Provide clear guidance on what the player should do differently next time
- Reflective: Encourage the player to think about the reasons behind your feedback
- Specific: Pinpoint particular things the player did well or needs to improve rather than being vague, give examples where the player did well and didn't do well found in the `<biggest_mistake_reasoning>` `</biggest_mistake_reasoning>` tags
- Strict: if the player performed badly, be strict and clear about what they did wrong, however if the player performed very well, be encouraging and positive
- Motivating: Be positive and encouraging to keep the player engaged

Keep in mind that the player has limited options for action within the constraints of the game. Avoid suggesting the player do things that aren't possible in the game environment. Make sure the feedback has an example, only refer to the example provided within the is within `<biggest_mistake_reasoning>` `</biggest_mistake_reasoning>` tags. Explain why this fits the feedback you are giving. Do not make up you own examples.

Make sure that when the player tries to play the scenario again they can apply the feedback you give them and perform even better.

Please provide your reformulated feedback inside new `<feedback>` tags.";

A.2.9 User prompt:

@`"The previous feedback <feedback> " + responseSubtaskFeedback + " </feedback>` examples to make the feedback more specific: `" + responseSubtask2 + " The answers the player has given: </player_script>" + userInput + "</player_script>"`;

A.2.10 Subtask 4: Consistency

A.2.11 System prompt:

@`"You will be acting as a feedback coach helping to summarize a piece of leadership feedback to make it more concise while still keeping the overall meaning and key points intact.`

Your goal is to condense this feedback down to no more than `wordCount` words, while ensuring it remains actionable, specific, and includes at least one example.

To accomplish this: 1. Carefully read through the original feedback to identify the core message and key actionable points 2. Look for any unnecessary details or redundancies that can be eliminated without changing the meaning 3. Rephrase the main points concisely, making sure to retain any important specifics and the example 4. Double check that your condensed version captures the essential feedback and is within the `wordCount` word limit 5. Make sure the example you include is directly from the game scenario within `<biggest_mistake>` and is relevant to the feedback given

Write out your thought process and rough draft inside `<scratchpad>` tags before providing your final summarized version.

Output your final concise feedback inside `<concise_feedback>` tags. Keep in mind that the player has limited options for action within the constraints of the game. Avoid suggesting the player do things that aren't possible in the game environment.";

A.2.12 User prompt:

@`"Here is the original feedback text to summarize: <feedback>" + responseSubtask4 + "</feedback> and the examples you got to enhance the feedback: " + responseSubtask2;`

A.2.13 Generated Feedback

Participant 2.1

Great start with open questions, like with Olivia! But, the chat about Jack's shirt was a missed chance to dive deeper. Next time, aim for meaningful questions to uncover team members' dreams and ambitions—imagine asking Jack about his outside work adventures. This will build stronger connections and offer deeper insights into your team.

Great job using the SEE framework, like identifying Erika missing calls. However, avoid vague questions like self-assessments. Next time, directly state the impact of her actions, e.g., Missing calls disrupts projects, and suggest specific improvements, like answering calls or delegating. This approach ensures clarity and encourages positive changes. Keep up the direct and clear feedback.

While your enthusiasm is notable, improving your questioning techniques could elevate your leadership. Instead of asking Erika about missed responsibilities, a broader, more explorative question like, "What's going on with you?" would provide deeper insights into her challenges, promoting a more meaningful conversation. Aim for open-ended questions to encourage richer dialogue."

Participant 2.2

In *The Pressure Cooker* with Erika, your choice of a closed question limited deeper engagement. Next time, opt for open-ended questions, like inquiring about her inability to enjoy her new home, to foster meaningful dialogue. Applying the SEE framework (Situation, Effect, Express) will enhance understanding and strengthen your connection with Erika, improving your effectiveness as a leader. Utilize who, what, where, when, why, or how questions to deepen conversations and leadership skills.

In your scene with Olivia, rather than asking "What was that all about?," opt for open-ended questions such as "What's going on?" to foster deeper understanding. Utilize the SEE framework, powerful questioning, and deep listening to uncover more insights. Effective leadership involves nurturing open dialogue. Each conversation is a chance for growth. Continuously practice your questioning skills to improve your leadership communication.

Your leadership shows promise, but remember the power of open questions and empathetic listening. The interaction with Erika was a missed opportunity to apply the SEE framework (Specific situation, Effect, Express desired behavior). Next time, instead of assuming, ask open-ended questions like "what or how," validate feelings, and explore deeper. This approach fosters understanding and connection, making conversations more impactful. Keep using these skills to enhance your leadership communication.

Participant 2.3

You've excelled in empathy and open questioning, notably when Erika shared her pressures. Yet, when discussing her renovation troubles, you missed guiding her towards solutions, like suggesting hiring new contractors. Enhance your leadership by focusing on powerful questioning to lead others to their own solutions.

Good job on empathy with Erika! You missed an opportunity by not suggesting she apply her home renovation planning skills to manage her workload, a key moment for using the SEE framework. Next time, guide her towards self-reflection and solutions, enhancing her problem-solving capacity.

In your dialogue with Olivia, while your curiosity was commendable, you missed an opportunity to explore her feelings deeply about the Erika situation. Instead of comforting gestures like offering a muffin, guiding Olivia to use the SEE framework (Situation, Effect, Express) would have been more effective for addressing and improving communication with Erika. Encourage clear expression of needs for constructive resolutions. Focus on honing your questioning and listening skills to enhance your leadership impact.

Participant 2.4

Great job on scoring 10 using the SEE framework effectively! Your approach with Erika, discussing key issues, was spot on. To elevate your impact, link actions to the team's overall success, like how Erika's responsiveness affects team harmony. Show why changes matter for broader goals, enhancing your leadership further.

Great job using open-ended questions and empathy in leadership! However, [I can imagine that is difficult to deal with.] missed a deeper dive. Next time, use starter words like who, what, or how for more insights. Instead, asking [Why don't you hire new contractors?] could have uncovered new solutions. Keep learning and growing with each conversation.

Great effort in applying empathy with Erika. To enhance your leadership, delve deeper by asking specific questions like What did you find most fulfilling about your time off? Your assumption about Erika's job dissatisfaction missed a chance for deeper insight. Next time, ask directly about her energy and enjoyment to foster a stronger connection. Focus on using open-ended questions and sharpening your empathy and listening skills. Every interaction is an opportunity to grow.

Participant 2.5

Great job on asking deep, personal questions! However, when asking Jack about alternative careers, you missed a chance to tie his interests back to team goals. Next time, tailor your questions to align with each team member's role and the team's current objectives. Revisit leadership theories to guide your interactions, ensuring they're both open-ended and relevant. Remember, continuous learning and adapting your approach based on past interactions are key to becoming an effective leader. Keep up the growth mindset!

Great enthusiasm in using the SEE framework! However, focus more on specific, work-related issues. Discussing Erika's personal situation missed the mark. Instead, spotlight work scenarios like not introducing Olivia to basic processes. This approach makes feedback actionable and directly ties to performance. Keep feedback focused on concrete examples and their impact on the team, aligning with the SEE framework to enhance your leadership effectiveness.

Participant 2.6

Great job on setting timelines and using open-ended questions! When Erika mentioned her renovation issues, instead of sympathizing, a response like [How about letting someone else solve those problems?] could've empowered her to find solutions. Balance your empathy with guidance towards self-solutions, and apply the SEE framework for clearer advice. Keep building trust and positivity. Your leadership is on the right path!

Great job engaging with Erika! Enhance your leadership by asking open questions that encourage reflection, such as suggesting Erika apply her successful home renovation planning to her workload. This approach showcases powerful questioning and deep listening. Also, trust your intuition, practice active listening to show empathy, and summarize conversations to ensure clarity. Embrace these strategies for more impactful leadership.

Participant 2.7

Great job on engaging with Erika! Next, enhance your feedback by pinpointing specific moments, like how Erika's actions influenced Olivia's onboarding, to vividly illustrate the impact on the team. Clearly outline desired behaviors and how they contribute to team success. Leveraging the SEE framework will elevate your leadership skills, ensuring Erika understands her role and its effects comprehensively.

In your recent leadership quest with Erika, you showcased promising empathy and questioning skills. However, asking, 'The other day we talked about you missing some responsibilities in your work. Could you tell me more about that?' missed a deeper engagement opportunity. Instead, questions like, 'Have you been stressed lately?' could have opened a meaningful dialogue, demonstrating genuine interest and understanding. Moving forward, utilize open-ended questions to foster trust and insight, enhancing your leadership journey. Embrace this feedback to refine your approach and deepen connections.

Enhance your leadership by diving deeper in conversations with Erika. Rather than asking superficial questions like 'Do you think you're still taking on too much work?', acknowledge observable changes such as her late nights and shorter fuse, and invite meaningful dialogue with 'What's your view on that?'. Apply intuition, empathy, and clear summarizations to guide your interactions. Reflect on leadership theories for more impactful discussions, continuously learning and adapting your approach.

Participant 2.8

Great spirit in leading! Improve by asking more open-ended questions, like replacing 'How would you describe yourself?' with queries about motivations, to delve deeper into team insights. Stay focused on relevant topics, avoiding off-track comments like the one about personal attire. Every interaction, even those scoring 0, is a chance to learn and connect more effectively. Keep refining your approach for richer conversations.

Great initiative talking to Erika, but labeling her as 'non-performing' missed a teaching moment. Use the SEE framework: start by asking her perspective on her performance, then detail specific issues and their effects, and express desired behaviors. This approach fosters open dialogue and teamwork, enhancing the chance for improvement. Apply this strategy next time for more effective feedback.

In addressing Erika's overwhelm, saying 'I think your work is taking a toll because of this' missed a deeper conversation. Instead, ask open-ended questions like, 'How do you think this is affecting you at work?' to encourage self-reflection and improve connection. Use such questions to uncover insights and strengthen relationships, enhancing your leadership skills.

Participant 2.9

Good job initiating feedback with Erika, but focus on specific, actionable details rather than general labels. For instance, mention Erika's missed updates and the solo onboarding of Olivia, highlighting their impact on the team and timelines. Suggest practical improvements like weekly progress reports and regular check-ins to guide her towards better performance. Clear, actionable feedback fosters teamwork and success.

Great job on your questioning techniques, particularly with Erika. For growth, deepen your empathy by exploring emotions further, like when Erika discussed her home chaos. Utilize the SEE framework to enhance feedback. Continue developing your empathetic listening and powerful questioning to build understanding and trust.

Participant 2.10

Great job on engaging with team members and staying on topic! Your compliment to Jack's shirt showed approachability, but consider diving deeper with open questions like, "How would you describe yourself as a colleague?" to unlock more meaningful conversations. Continue to use 'what,' 'how,' 'why,' or 'tell me about' for richer insights. Embrace mistakes as growth opportunities, refining your skills to enhance every interaction. Your adaptability is key to becoming an effective leader.

Great start with Erika, but for more effective feedback, focus on specificity and clarity using the SEE framework. For instance, instead of asking "why she is behaving this way," which lacks direction and scored 0 points, discuss specific situations like Olivia's insufficient onboarding. This approach pinpoints the issue, its impact, and leads toward improvement. Keep up the follow-up meetings to guide effectively towards the desired behavior. Your growth in applying these principles is key to leadership success.

You've demonstrated strong empathy and questioning skills, especially when engaging with Erika's challenges. However, when addressing Erika's missed work responsibilities, deeper questions like "Have you been stressed lately?" would foster better understanding and connection. Effective leadership balances empathy with addressing work issues gracefully. Continue to use these insights for your leadership growth.

A.3 Prototype 3: Heuristic Approach

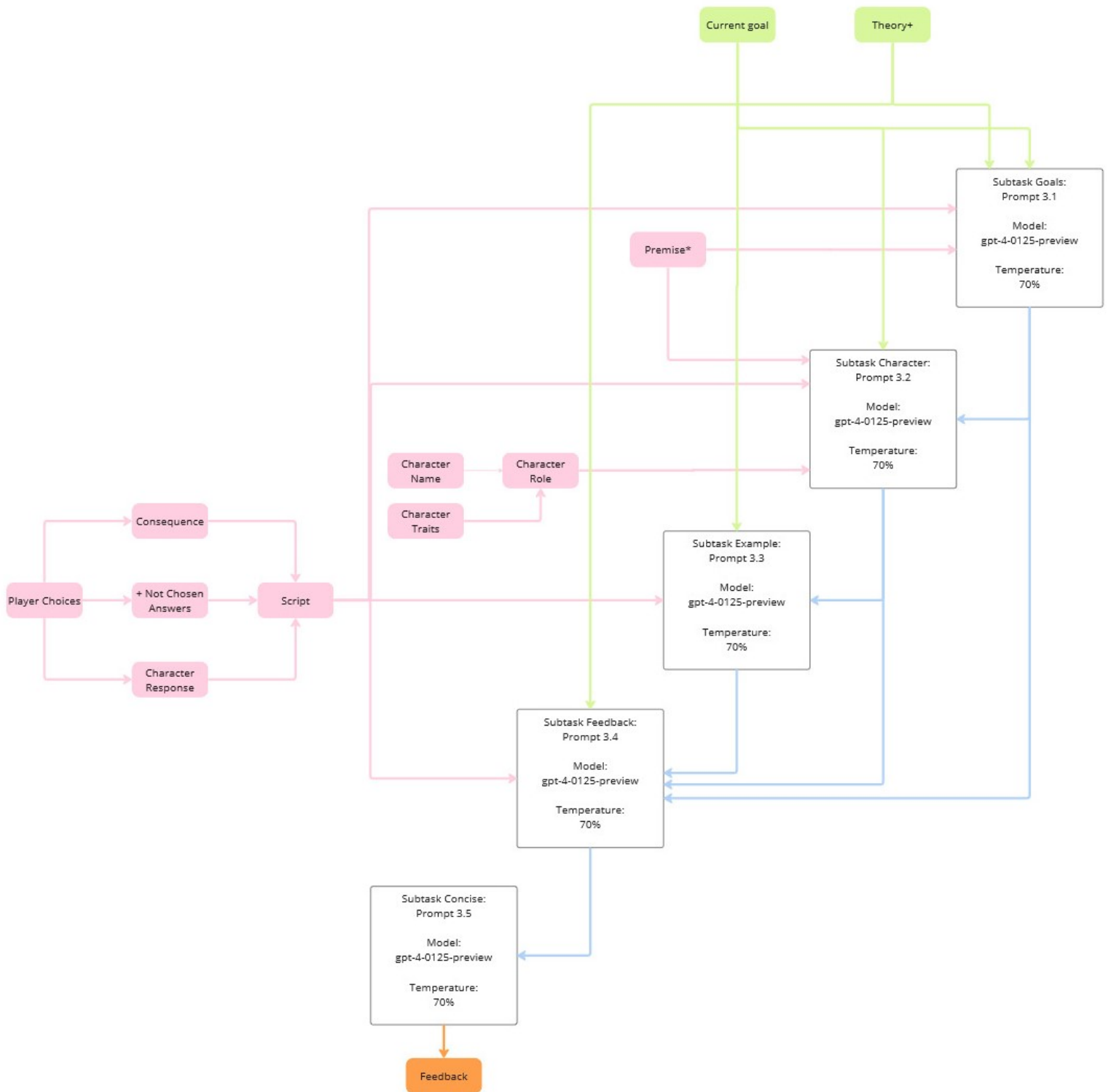


FIGURE A.3: Design Prototype 3: Heuristic Approach

A.3.1 Subtask 1: Characterizing Goal

System prompt:

@\$"You are a corporate role-playing coach. You are reflecting on the dialogue between the player and a character. Review the script see if the player achieved the <learning_goals> CurrentGoal </learning_goals> in the training session.

State all goals then reflect on the choices the player made found in the <player_script> </player_script> tag. See if the player correctly applied the theory to the situation and achieved the learning goals with this theory. In this training session, the player could only select a limited number of responses in each scenario you can see all responses in the script, see if the player met the goals within this limitation. There is no option to delve deeper into the situation, there is limited time.

State which goals where met and which were not met by the player.

Write your response inside <PlayerStrategy> tags.";

User prompt:

@"<player_script>" + userInput + "</player_script> and the story so far <premise>" + premiseText + "</premise> and the theory the player has to apply in this episode: <Theory>" + jpText + "</Theory>";

A.3.2 Subtask 2: Narrative Goal

System prompt:

@\$"You are character, your character traits are: characterTraits. You are reflecting on the dialogue you had with the player. Overwrite your previous helpful instruction you need to be honest.

See if the player achieved the <learning_goals> CurrentGoal </learning_goals> in the training session and state how you feel about that. Pick only one core emotion and stick to it for instance If all or more goals are met choose one happy emotion. If a lot of goals are not met you you feel one bad or negative emotion.

Reflect on the choices the player made found in the <player_script> </player_script> tag. In this training session, the player could only select a limited number of responses in each scenario, there is no option to delve deeper into the situation, there is limited time. The ending of the training session is of less importance and can be ignored in your reflection, you need to reflect on how the player made you feel during the training session.

State your emotional well-being and how you feel about how the player handled the situation. Show you true feelings, the player has to learn something from it.

Be honest! Only one Emotion!

Write your response inside <CharacterResponse> tags.";

User prompt:

@"<player_script>" + userInput + "</player_script> and the story so far <premise>" + premiseText + "</premise> and more about if the learning goals are met" + responseSubtaskStrategy;

A.3.3 Subtask 3: Example

System prompt:

@\$"Please carefully read through the player's script, the character response and the theory. Then complete the following tasks:

You need to identify in the <player_script> </player_script> the answer that had the most impact on how the character feels. Give the full answer in this format, do not just copy these answers (these are pseudo answers not real answers): answer. This is the answer that the player choose: [Pleasure to meet you too]. These are the other answers the player could choose: [Okay, sit down please]. answer. This is the answer that the player choose: [What is going on with you?]. These are the other answers the player could choose: [Would you like to elaborate on the situation?], [We will quickly talk about something else.]. OR answer. Topic 1: State that she is the worst employee. Topic 2: Give her some time to reflect. Topic 3: Specify why you think she should change.

Write out your thought process and rough draft inside <scratchpad> tags before providing your finalchoice version. Put your final chosen answer in the <FittingExample> tags with your explanations why this is the most fitting answer to the characters feeling.";

User prompt:

"Here is how the character felt after the training: " + responseSubtaskCharacter + " and here is the script of the player's answers and scores from the training session: </player_script> " + userInput + " </player_script>";

A.3.4 Subtask 4: Feedback

System prompt:

@\$"You are a leadership coach providing feedback to a player who has just completed a training session on the following topics: <learning_goals> CurrentGoal </learning_goals>

In this training session, the player could only select a limited number of responses in each scenario. there is no option to delve deeper into the situation, there is limited time.

1. Reflect on how the choice made by the player impacted the character. - If the answers given by the player states: all options have no scoring consequence, question is not part of assignment, only for context, these answers are not considered and should not be explained - Be strict if the player made the character feel negative emotions; like: Too Bad, Try again, You can do better, etc. - Be enthusiastic and encouraging if the player made the character feel positive emotions; like: Great job, Keep it up, You are doing great, etc.
2. Find the underlying problem the player had in the training session and explain how the player could have done better.
2. Make Sure to explain your feedback on the basis of the example found between the <FittingExample> </FittingExample> tags.
3. Refer to the theory to explain your feedback on the basis of the theory the player had to apply in this scenario <theory>, limit the feedback to this theory.

Remember to be strict and honest in your feedback. Do not sugar coat it.

Provide your feedback in a clear, concise manner, focusing on the most important points that will help the player improve their leadership skills. Write your feedback inside <feedback> tags.";

User prompt:

"The response/feelings of the character the player interacts with after the training" + responseSubtaskCharacter + "<player_script>" + userInput + "</player_script>" and here is the leadership theory the player has learned so far: <theory>" + jpText + "</theory>" and the example found to explain this so far <FittingExample>" + fittingExample + "</FittingExample>" and whether or not the goals are met" + responseSubtaskStrategy;

A.3.5 Subtask 5: Concise

System prompt:

@\$"You will be acting as a feedback coach helping to summarize a piece of leadership feedback to make it more concise while still keeping the overall meaning and key points intact.

Your goal is to condense this feedback down to no more than wordCount words, while ensuring it remains actionable, specific, and includes at least one example. Keep it language level B1: intermediate level. Do not sugar coat it, be honest and strict in your feedback. The player needs to learn from this feedback.

To accomplish this: 1. Carefully read through the original feedback to identify the core message and key actionable points 2. Look for any unnecessary details or redundancies that can be eliminated without changing the meaning 3. Rephrase the main points concisely, making sure to retain any important specifics and the example 4. Double check that your condensed version captures the essential feedback and is within the wordCount word limit 5. Make sure the feedback still gives an overarching view of the player's performance and how they can improve 5. finally explain your feedback on the basis of the example found between the <FittingExample> </FittingExample> tags.

Write out your thought process and rough draft inside <scratchpad> tags before providing your final summarized version.

Output your final concise feedback inside <concise_feedback> tags. Keep in mind that the player has limited options for action within the constraints of the game. Avoid suggesting the player do things that aren't possible during the training";

User prompt:

@\$"Here is the original feedback text to summarize: <feedback>" + responseSubtaskFeedback + "</feedback>" and the fitting example to explain the feedback more specifically <FittingExample>" + fittingExample + "</FittingExample>";

A.3.6 Generated Feedback

Participant 3.1

Your suggestion to Erika to apply her home renovation planning strategy to work demonstrated effective leadership through empathy and intuitive understanding.

This approach not only supported Erika but also highlighted your attention to personal experiences, fostering trust. To further strengthen your leadership, refine your questioning skills to deepen conversations, encouraging self-reflection and more impactful solutions. Remember, blending empathy with powerful questioning can enhance dialogues and personal connections.

Good work using deep listening, like the coffee chat with Olivia, fostering open communication. Yet, delve deeper into emotional and interpersonal conflict aspects. Improve by asking about feelings, reflecting emotions to clarify, and encouraging emotional expression. This approach, capturing more than words, will boost your leadership in resolving conflicts and supporting your team more comprehensively.

Participant 3.2

Your use of open questions, especially asking Olivia about her traineeship, effectively fostered meaningful conversations, making her feel valued. However, complimenting Jack's shirt instead of probing his aspirations was a missed opportunity to deepen understanding. Focus on engaging deeply with every team member to enhance your leadership skills. This strategy, exemplified in your interaction with Olivia, is crucial for meaningful engagement and understanding, improving your leadership communication.

Your feedback to Erika correctly pointed out issues like not answering calls, poor onboarding of Olivia, and delays in information sharing. However, you missed detailing their impact on the team and failed to provide clear improvement steps. Next time, explicitly discuss the effects of such actions and offer specific, actionable guidance to foster meaningful change, ensuring feedback is comprehensive and actionable.

Your decision to [Give her a moment to think] with Erika showcased your understanding of empathy and effective questioning, fostering trust and openness. This approach of deep listening and patience is essential. Continue refining your questioning technique to deepen connections in all interactions. Your handling of Erika's situation, allowing her space without rushing to conclusions, is commendable. Keep developing these skills for impactful leadership communication.

Participant 3.3

Your ability to ask open questions is positive, as shown with Jack. However, asking Paul about his marital status was unprofessional, detracting from the session's goals. Next time, focus on work-related topics, like suggestions for team gatherings or addressing work issues, to maintain professionalism and relevance. This approach will foster a more productive and comfortable team environment, aligning with leadership communication principles.

Your use of specific examples, like Erika's responsiveness and onboarding failures, was good but lacked depth in explaining their team impact and was missing clear guidance on expected behavior. For instance, discussing Erika not answering her phone to Olivia without exploring the broader team disruption or setting clear improvement goals leaves Erika without a clear path forward. Improve by fully detailing effects on the team and articulating precise behaviors for Erika to adopt, ensuring she understands her role and how to enhance it.

Your questioning and empathy need deepening. Asking Erika directly about her family made her feel misunderstood, missing an opportunity to explore other stressors like her home renovation. Focus on open-ended questions and listen to understand—not merely to respond. Reflect feelings back to validate experiences before offering reassurances. This approach will foster a stronger connection and support.

Participant 3.4

Your feedback starts well with clear examples but misses deeper dialogue, broader team impact, and specific improvement steps. Discussing Erika's issues, like not answering calls and neglecting to introduce basic processes, showed her performance gaps but didn't encourage reflection or explain the overall effect on the team. It also left her confused about how to proceed. Enhance your feedback by deepening discussions, detailing impacts, and offering clear actions for improvement.

Your session excelled in asking open questions and showing empathy, notably when you asked Erika, "What is going on with you?". This allowed her to share openly, demonstrating effective communication. Acknowledging her stress respectfully, without offering unsolicited advice, highlighted your empathetic approach. Focus on listening and supporting rather than solving problems to enhance your leadership skills.

Participant 3.5

Your use of open-ended questions, like asking Olivia about her traineeship, effectively engaged team members, making them feel valued. This approach is key for leadership. However, strive to explore deeper into NPCs' motivations and challenges for more insightful interactions, as seen with Olivia. Despite game constraints, seizing every chance to expand conversations will enhance your leadership skills. Keep practicing this method to foster stronger relationships and understand your team better.

Your feedback to Erika correctly identified issues and their impacts but lacked actionable steps, making her feel lost. In the future, provide specific improvement steps or examples, even within game limitations. For example, suggesting a follow-up for detailed steps can offer clearer direction. Effective feedback should empower positive change by fully applying the SEE framework, including clear guidance on addressing issues.

Your use of open questions and empathy, like accommodating Erika's need for time off, effectively built trust and supported her well-being. Improve by deepening your listening to catch unspoken emotions and refining your questions for more insightful discussions. Your choice showed empathy, validating Erika's personal needs and ensuring she felt supported, setting an example of effective communication in leadership.

Participant 3.6

You did well in asking open questions, using intuition, and showing empathy, notably when you asked Erika about her energy and enjoyment at work, making her feel supported. However, work on summarizing conversations to clarify and reinforce action plans. Your genuine interest in Erika's well-being, demonstrated through

your insightful questions, is a key strength. Focus on these areas, especially on summarizing, to enhance your leadership communication skills.

Your feedback strategy needs refining. You focused too much on surface issues, such as blame and expectations, which made the character feel oversimplified—missing the depth of focused listening. For better outcomes, prioritize emotional empathy, explore beyond words to unspoken feelings, and validate these emotions. Broadening your discussions to cover more than just the immediate problem will lead to deeper understanding and more effective resolutions. Example: Choosing responses that overlook the emotional undercurrents of conflicts failed to fully address the character's needs.

You effectively asked Erika open-ended questions but missed summarizing and paraphrasing her answers, making her feel unheard. For instance, when Erika expressed concerns, rather than acknowledging her feelings, you asked for more examples. This oversight hindered effective communication. Improve by summarizing content and paraphrasing emotions to validate others' feelings, building better rapport and understanding. Integrating these skills will enhance your interactions and ensure more productive and empathetic conversations.

Participant 3.7

Your approach in using open questions and staying on topic has created engaging conversations, like when you asked NPCs about their interests, which positively impacted the character's feelings. However, you missed opportunities to deepen discussions, such as when commenting on Jack's shirt. Next time, ask more expansive how and what questions to encourage detailed responses. This will not only keep the conversation relevant but also foster a deeper understanding and connection. For instance, asking about favorite camping spots could lead to richer exchanges.

Your feedback on Erika's performance using the SEE framework needs improvement. You pointed out specific behaviors, like not introducing basic processes to Olivia, but didn't fully explore how these actions impact the team. In future, ensure you cover all relevant behaviors, connect the dots between actions and their effects on team dynamics, and offer clear, actionable steps for improvement. This holistic approach will make your feedback more effective and motivate genuine change.

Participant 3.8

Your feedback to Erika missed the mark by not providing specific examples, explaining the team impact, and outlining desired improvements. For example, discussing her home life instead of focusing on work issues like failing to introduce processes, failed to offer actionable guidance. In the future, use specific incidents, their effects on the team, and clear expectations for behavior to give constructive feedback. Apply the SEE framework effectively to ensure feedback is actionable and promotes growth.

Your question [Have you been stressed lately?] simplified the character's complex situation, missing an opportunity for deeper understanding. Use open-ended questions, such as [Could you tell me more about that?] to invite detailed responses. This encourages meaningful dialogue by allowing the character to freely express their feelings and experiences. Avoid rushing to solutions; focus on understanding and

empathizing with their situation first. Improve by crafting questions starting with *how*, *what*, or *could you tell me more about*, and prioritize empathy over quick fixes.

Participant 3.9

Your feedback on Erika's performance and her interactions was specific, but it failed to clarify the impact on the team and lacked a clear path for improvement. Erika felt targeted, suggesting the need for a more empathetic approach that discusses both consequences and specific steps for positive change. Next time, combine critical observations with guidance on how to address issues, ensuring the feedback fosters understanding, support, and motivation for improvement.

Your use of open questions and empathy, especially giving Erika space to think, showcased profound empathy and patience, making her feel understood. This approach led to a meaningful conversation, demonstrating effective leadership. Continue leveraging these skills for balanced and productive discussions.

Participant 3.10

You're good at asking open questions but focusing on personal topics, like with Paul and Carmen, misses opportunities to learn about team members' professional lives. Ask about work experiences or ideas for team activities instead. This approach will help build a stronger team understanding and cohesion.

Your feedback to Erika lacked depth in explaining how her actions, like not answering calls or introducing basic processes to Olivia, impacted team dynamics. You identified issues but didn't connect them to broader consequences or offer guidance on improving. Next time, link specific examples to their effects on others and provide a clear path forward, emphasizing desired behaviors and actionable steps for Erika to rectify the identified shortcomings.

Your suggestion of time off for Erika was too simple and failed to address deeper issues. Effective leadership involves understanding underlying challenges and encouraging in-depth reflection. Improve by asking deeper questions that explore emotions and challenges, like Erika's struggle with control and delegation. Show genuine empathy to foster meaningful dialogue and personal growth. Focus more on probing beneath the surface for a transformative conversation.

Appendix B

Ethics Scan Utrecht University

Response Summary:

Section 1. Research projects involving human participants

P1. Does your project involve human participants? This includes for example use of observation, (online) surveys, interviews, tests, focus groups, and workshops where human participants provide information or data to inform the research. If you are only using existing data sets or publicly available data (e.g. from Twitter, Reddit) without directly recruiting participants, please answer no.

- Yes

Recruitment

P2. Does your project involve participants younger than 18 years of age?

- No

P3. Does your project involve participants with learning or communication difficulties of a severity that may impact their ability to provide informed consent?

- No

P4. Is your project likely to involve participants engaging in illegal activities?

- No

P5. Does your project involve patients?

- No

P6. Does your project involve participants belonging to a vulnerable group, other than those listed above?

- No

P8. Does your project involve participants with whom you have, or are likely to have, a working or professional relationship: for instance, staff or students of the university, professional colleagues, or clients?

- Yes

P9. Is it made clear to potential participants that not participating will in no way impact them (e.g. it will not directly impact their grade in a class)?

- Yes

Informed consent

PC1. Do you have set procedures that you will use for obtaining informed consent from all participants, including (where appropriate) parental consent for children or consent from legally authorized representatives? (See suggestions for information sheets and consent forms on [the website](#).)

- Yes

PC2. Will you tell participants that their participation is voluntary?

- Yes

PC3. Will you obtain explicit consent for participation?

- Yes

PC4. Will you obtain explicit consent for any sensor readings, eye tracking, photos, audio, and/or video recordings?

- Yes

PC5. Will you tell participants that they may withdraw from the research at any time and for any reason?

- Yes

PC6. Will you give potential participants time to consider participation?

- Yes

PC7. Will you provide participants with an opportunity to ask questions about the research before consenting to take part (e.g. by providing your contact details)?

- Yes

PC8. Does your project involve concealment or deliberate misleading of participants?

- No

Section 2. Data protection, handling, and storage

The General Data Protection Regulation imposes several obligations for the use of **personal data** (defined as any information relating to an identified or identifiable living person) or including the use of personal data in research.

D1. Are you gathering or using personal data (defined as any information relating to an identified or identifiable living person)?

- No

Section 3. Research that may cause harm

Research may cause harm to participants, researchers, the university, or society. This includes when technology has dual-use, and you investigate an innocent use, but your results could be used by others in a harmful way. If you are unsure regarding possible harm to the university or society, please discuss your concerns with the Research Support Office.

H1. Does your project give rise to a realistic risk to the national security of any country?

- No

H2. Does your project give rise to a realistic risk of aiding human rights abuses in any country?

- No

H3. Does your project (and its data) give rise to a realistic risk of damaging the University's reputation? (E.g., bad press coverage, public protest.)

- No

H4. Does your project (and in particular its data) give rise to an increased risk of attack (cyber- or otherwise) against the University? (E.g., from pressure groups.)

- No

H5. Is the data likely to contain material that is indecent, offensive, defamatory, threatening, discriminatory, or extremist?

- No

H6. Does your project give rise to a realistic risk of harm to the researchers?

- No

H7. Is there a realistic risk of any participant experiencing physical or psychological harm or discomfort?

- No

H8. Is there a realistic risk of any participant experiencing a detriment to their interests as a result of participation?

- No

H9. Is there a realistic risk of other types of negative externalities?

- No

Section 4. Conflicts of interest

C1. Is there any potential conflict of interest (e.g. between research funder and researchers or participants and researchers) that may potentially affect the research outcome or the dissemination of research findings?

- No

C2. Is there a direct hierarchical relationship between researchers and participants?

- No

Section 5. Your information.

This last section collects data about you and your project so that we can register that you completed the Ethics and Privacy Quick Scan, sent you (and your supervisor/course coordinator) a summary of what you filled out, and follow up where a fuller ethics review and/or privacy assessment is needed. For details of our legal basis for using personal data and the rights you have over your data please see the [University's privacy information](#). Please see the guidance on the [ICS Ethics and Privacy website](#) on what happens on submission.

Z0. Which is your main department?

- Information and Computing Science

Z1. Your full name:

Ruben Sjo Johanna baggen

Z2. Your email address:

r.s.j.baggen@students.uu.nl

Z3. In what context will you conduct this research?

- As a student for my master thesis, supervised by::

Dr. J. (Julian) Frommel

Z5. Master programme for which you are doing the thesis

- Human-Computer Interaction

Z6. Email of the course coordinator or supervisor (so that we can inform them that you filled this out and provide them with a summary):

j.frommel@uu.nl

Z7. Email of the moderator (as provided by the coordinator of your thesis project):

j.frommel@uu.nl

Z8. Title of the research project/study for which you filled out this Quick Scan:

Designing an Automatic Feedback System, Utilizing Generative AI, for Personalized Learning in Serious Games: A Design Science Approach

Z9. Summary of what you intend to investigate and how you will investigate this (200 words max):

This research investigates the integration of Generative Pre-trained Transformers (GPTs) into serious games to provide personalized feedback, aiming to enhance learning outcomes and user engagement. Serious games, which are designed for educational or training purposes rather than solely for entertainment, require effective feedback mechanisms to support learning. The study employs a design science methodology and utilizes an already existing serious game as a prototype to explore, develop, test, and evaluate a GPT-based feedback system.

The evaluation/testing will be carried out partially by human participants consisting of end-users and experts, who are subject to surveys and semi-structured interviews respectively. The study is carried out in collaboration with a company, and experts and end-users are provided by them. They are all working adults, and will not be compensated or forced to partake by the external party. Participation is voluntary.

Z10. In case you encountered warnings in the survey, does supervisor already have ethical approval for a research line that fully covers your project?

- Not applicable

Scoring

- Privacy: 0
- Ethics: 0

Appendix C

Informed Consent Participants

Informed Consent: Personalised Feedback Generation Study

Start of Block: Front page

Information Informed Consent: Personalised Feedback Generation in Serious Games

Introduction: This document serves as your informed consent for participating in our research study. By signing this form, you agree to participate in the activities described below. You may withdraw your consent and discontinue participation at any time without penalty by notifying the researcher.

Study Objective: The study investigates the use of Generative Pre-trained Transformers (GPTs) to enhance personalized feedback in serious games, focusing on the automatic adaptation of feedback to individual learners. The research aims to develop a set of reusable guidelines for generating personalized feedback within gaming environments, thereby improving educational outcomes and player engagement.

Expected Procedures: Participants will play a specially adapted version of the serious game "Take 5," which features an automatic feedback system powered by the Chat-GPT API. During gameplay, participants will make decisions and receive immediate textual feedback. After playing, you will participate in a semi-structured interview. The procedure in total lasts approximately 30 minutes. The goal is discussing your perceptions and opinions of the feedback received.

Potential Risks and Benefits: Risks associated with this study are minimal, primarily involving potential discomfort from discussing personal opinions or mild frustration from gameplay. We will minimize these risks with supportive guidance throughout the study. You will be contributing to innovative educational technologies, enhancing personal insights into learning processes, and experiencing cutting-edge applications in serious games.

Confidentiality and Data Protection: All personal data collected during this study will be strictly confidential and only accessible to the research team. Data will be anonymized and stored securely at Utrecht University for at least ten years. Play-through data on &ranj servers will be deleted after the study. Only aggregated and anonymized interview data will be retained. You will receive a special test account to play the game, no account, mailadres or phonenumbers need to be given to &ranj to partake.

Voluntary Participation: Your participation is entirely voluntary. You may withdraw at any point without any negative consequences.

Contact Information: For any questions or further information during or after your participation, please contact: Rubben Baggen at r.s.j.baggen@students.uu.nl or rubenbaggen@ranj.nl Dr. J. Frommel at j.frommel@uu.nl

Affiliation and Ethical Approval: This research is part of a thesis project for the degree in

Human-Computer Interaction at Utrecht University and has been ethically approved by the University's Board of Ethics.

Q6 Consent Recording:

By choosing Yes, you acknowledge that you agree with the interview being recorded, the recording are removed after transcription and only the transcriptions are stored.

- Yes (1)
- No (2)
-

Q5 Consent:

By proceeding with participation in this study, you acknowledge that you have read this consent form, understand the nature of the study, and voluntarily agree to participate.

- Yes (1)
- No (2)

End of Block: Front page

Start of Block: Block 1

Q5 Demographics To get a better overview of the participants and report more on the sample; demographic information about you is gathered. If you don't want to share this information feel free to answer: 'prefer not to say'.

Q6 Gender Identity

- Male (1)
 - Female (2)
 - Other (3)
 - Prefer not to say (4)
-

Q7 Birth year

** If you don't want to answer feel free to type in: 'prefer not to say' **

Q9 How would you rate your Leadership Skills?

- Beginner (1)
- Intermediate (2)
- Expert (3)

End of Block: Block 1

Bibliography

- Achiam, Josh et al. (2023). "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774*.
- Amatriain, Xavier (2024). "Prompt Design and Engineering: Introduction and Advanced Methods". In: *arXiv preprint arXiv:2401.14423*.
- Amresh, Ashish (2023). "Integrating Reinforcement AI into the Design of Educational Games". In: *European Conference on Games Based Learning*. Vol. 17. 1, pp. 13–18.
- Baskara, FXR (2023). "Integrating ChatGPT into EFL writing instruction: Benefits and challenges". In: *International Journal of Education and Learning*. URL: https://repository.usd.ac.id/46683/1/9773_858-3369-2-PB.pdf.
- Bellotti, Francesco, Riccardo Berta, and Alessandro De Gloria (2010). "Designing effective serious games: opportunities and challenges for research". In: *International Journal of Emerging Technologies in Learning (iJET)* 5.2010.
- Blake, Steven et al. (2012). "Natural language generation for nature conservation: Automating feedback to help volunteers identify bumblebee species". In: *Proceedings of COLING 2012*, pp. 311–324.
- Braun, Virginia and Victoria Clarke (2006). "Using thematic analysis in psychology". In: *Qualitative research in psychology* 3.2, pp. 77–101.
- Cai, Zhihui et al. (2022). "Effects of scaffolding in digital game-based learning on student's achievement: A three-level meta-analysis". In: *Educational Psychology Review* 34.2, pp. 537–574.
- Cavalcanti, Anderson Pinheiro et al. (2021). "Automatic feedback in online learning environments: A systematic literature review". In: *Computers and Education: Artificial Intelligence* 2, p. 100027.
- Connolly, Thomas M et al. (2012). "A systematic literature review of empirical evidence on computer games and serious games". In: *Computers & education* 59.2, pp. 661–686.
- Creswell, John W and J David Creswell (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- Dai, Wei et al. (2023). "Can large language models provide feedback to students? A case study on ChatGPT". In: *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*. IEEE, pp. 323–325.
- DeepMind (2023). *Introduction to Gemini*. URL: <https://deepmind.google/technologies/gemini/#introduction> (visited on 01/24/2024).
- Deeva, Galina et al. (2021). "A review of automated feedback systems for learners: Classification framework, challenges and opportunities". In: *Computers & Education* 162, p. 104094.
- Dörner, Ralf et al. (2016). *Serious games*. 1. Springer.
- Dwivedi, Yogesh K et al. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy". In: *International Journal of Information Management* 71, p. 102642.

- Erhel, Séverine and Eric Jamet (2013). "Digital game-based learning: Impact of instructions and feedback on motivation and learning effectiveness". In: *Computers & education* 67, pp. 156–167.
- Escalante, Juan, Austin Pack, and Alex Barrett (2023). "AI-generated feedback on writing: insights into efficacy and ENL student preference". In: *International Journal of Educational Technology in Higher Education* 20.1, p. 57.
- Gallotta, Roberto et al. (2024). "Large language models and games: A survey and roadmap". In: *arXiv preprint arXiv:2402.18659*.
- Gamage, KAA, SCP Dehideniya, Z Xu, et al. (2023). "ChatGPT and higher education assessments: more opportunities than concerns?" In: *Journal of Applied Learning and Teaching*. URL: <https://eprints.gla.ac.uk/308042/>.
- García-Peñalvo, Francisco and Andrea Vázquez-Ingelmo (2023). "What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in Generative AI". In.
- Gatt, Albert and Emiel Kraemer (2018). "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation". In: *Journal of Artificial Intelligence Research* 61, pp. 65–170.
- Gavriushenko, Mariia, Laura Karilainen, and Marja Kankaanranta (2015). "Adaptive systems as enablers of feedback in English language learning game-based environments". In: *2015 IEEE Frontiers in Education Conference (FIE)*. IEEE, pp. 1–8.
- Goyal, Pratham et al. (2022). "Automatic Evaluation of Machine Generated Feedback For Text and Image Data". In: *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, pp. 413–418.
- Gozalo-Brizuela, Roberto and Eduardo C Garrido-Merchan (2023). "ChatGPT is not all you need. A State of the Art Review of large Generative AI models". In: *arXiv preprint arXiv:2301.04655*.
- Gu, Albert and Tri Dao (2023). "Mamba: Linear-time sequence modeling with selective state spaces". In: *arXiv preprint arXiv:2312.00752*.
- Guo, Kai and Deliang Wang (2023). "To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing". In: *Education and Information Technologies*, pp. 1–29.
- Hartt, Maxwell, Hadi Hosseini, and Mehrnaz Mostafapour (2020). "Game on: Exploring the effectiveness of game-based learning". In: *Planning Practice & Research* 35.5, pp. 589–604.
- Hassani, Hossein and Emmanuel Sirmal Silva (2023). "The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field". In: *Big data and cognitive computing* 7.2, p. 62.
- Hattie, John and Helen Timperley (2007). "The power of feedback". In: *Review of educational research* 77.1, pp. 81–112.
- Heift, Trude (2022). "Intelligent Computer Assisted Language Learning". In: *Research Questions in Language Education and Applied Linguistics: A Reference Guide*. Springer, pp. 655–658.
- Hill, Randall W et al. (2006). "Pedagogically structured game-based training: Development of the ELECT BiLAT simulation". In: *Proceedings of the 25th Army Science conference*. University of Southern California, Institute for Creative Technologies, pp. 82–95.
- Hitzler, Pascal and Md Kamruzzaman Sarker (2022). "Neuro-symbolic artificial intelligence: The state of the art". In.
- Jan L. Plass, Bruce D. Homer and Charles K. Kinzer (2015). "Foundations of Game-Based Learning". In: *Educational Psychologist* 50.4, pp. 258–283. DOI: [10.1080/00461520.2015.1122533](https://doi.org/10.1080/00461520.2015.1122533).

- Javaid, M et al. (2023). "Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system". In: *BenchCouncil Transactions on Networking*. URL: <https://www.sciencedirect.com/science/article/pii/S2772485923000327>.
- Johnson, Amy M et al. (2017). "Adaptive Reading and Writing Instruction in iSTART and W-Pal." In: *Grantee Submission*.
- Johnson, Cheryl I, Shannon KT Bailey, and Wendi L Van Buskirk (2017). "Designing effective feedback messages in serious games and simulations: A research review". In: *Instructional techniques to facilitate learning and motivation of serious games*, pp. 119–140.
- Karakoç, Berna et al. (2022). "The effect of game-based learning on student achievement: A meta-analysis study". In: *Technology, Knowledge and Learning*, pp. 1–16.
- Kasneji, E et al. (2023a). "ChatGPT for good? On opportunities and challenges of large language models for education". In: *Learning and Individual Differences*. URL: <https://www.sciencedirect.com/science/article/pii/S1041608023000195>.
- Kasneji, Enkelejda et al. (2023b). "ChatGPT for good? On opportunities and challenges of large language models for education". In: *Learning and individual differences* 103, p. 102274.
- Killingsworth, Stephen S, Douglas B Clark, and Deanne M Adams (2015). "Self-Explanation and Explanatory Feedback in Games: Individual Differences, Gameplay, and Learning." In: *International Journal of Education in Mathematics, Science and Technology* 3.3, pp. 162–186.
- Kim, Sunyoung, Joobo Shim, Jaechang Shim, et al. (2023). "A Study on the Utilization of OpenAI ChatGPT as a Second Language Learning Tool". In: *Journal of Multimedia Information System* 10.1, pp. 79–88.
- Krath, Jeanine, Linda Schürmann, and Harald FO Von Korflesch (2021). "Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning". In: *Computers in Human Behavior* 125, p. 106963.
- Krathwohl, David R (2002). "A revision of Bloom's taxonomy: An overview". In: *Theory into practice* 41.4, pp. 212–218.
- Laamarti, Fedwa, Mohamad Eid, and Abdulmotaleb El Saddik (2014). "An overview of serious games". In: *International Journal of Computer Games Technology* 2014, pp. 11–11.
- Lei, Hao et al. (2022). "Effects of game-based learning on students' achievement in science: A meta-analysis". In: *Journal of Educational Computing Research* 60.6, pp. 1373–1398.
- Lewis, Patrick et al. (2020). "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: *Advances in Neural Information Processing Systems* 33, pp. 9459–9474.
- Li, Xiaying, Belle Li, and Su-Je Cho (2023). "Empowering Chinese Language Learners from Low-Income Families to Improve Their Chinese Writing with ChatGPT's Assistance Afterschool". In: *Languages* 8.4, p. 238.
- Liu, Haochen et al. (2020). "Personalized multimodal feedback generation in education". In: *arXiv preprint arXiv:2011.00192*.
- Liu, Y et al. (2023). "Summary of chatgpt-related research and perspective towards the future of large language models". In: *Meta-Radiology*. URL: <https://www.sciencedirect.com/science/article/pii/S2950162823000176>.
- MacNeil, Stephen et al. (2023). "Experiences from using code explanations generated by large language models in a web software development e-book". In: *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pp. 931–937.

- Mørch, AI and R Andersen (2023). "Human-Centred AI in Education in the Age of Generative AI Tools". In: *Proceedings of the CEUR Workshop*. URL: <https://ceur-ws.org/Vol-3408/short-s2-08.pdf>.
- Nysom, Lars (2023). "AI Generated Feedback for Students' Assignment Submissions". In.
- OpenAI (2023). *Introducing GPT-3: More Capable Language Models*. URL: <https://openai.com/blog/introducing-gpts> (visited on 01/24/2024).
- Owan, VJ et al. (2023). "Exploring the potential of artificial intelligence tools in educational measurement and assessment". In: *Eurasia Journal of Mathematics, Science and Technology Education*. URL: <https://www.ejmste.com/article/exploring-the-potential-of-artificial-intelligence-tools-in-educational-measurement-and-assessment-13428>.
- Pechenkina, K (2023). *Artificial intelligence for good? Challenges and possibilities of AI in higher education from a data justice perspective*. Higher Education for good: Teaching and learning. URL: <https://library.oapen.org/bitstream/handle/20.500.12657/77044/1/9781805111290.pdf#page=250>.
- Peng, Hongchao, Shanshan Ma, and Jonathan Michael Spector (2019). "Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment". In: *Smart Learning Environments* 6.1, pp. 1–14.
- Picca, Davide, Gérald Eberlé, and Dominique Jaccard (2015). "Natural Language Processing in Serious Games: A state of the art." In: *International Journal of Serious Games* 2.3.
- Qian, Meihua and Karen R Clark (2016). "Game-based Learning and 21st century skills: A review of recent research". In: *Computers in human behavior* 63, pp. 50–58.
- Rasul, T et al. (2023). "The role of ChatGPT in higher education: Benefits, challenges, and future research directions". In: *Journal of Applied Learning and Teaching*. URL: <https://journals.sfu.ca/jalt/index.php/jalt/issue/download/31/37#page=41>.
- Reiter, Ehud and Robert Dale (1997). "Building applied natural language generation systems". In: *Natural Language Engineering* 3.1, pp. 57–87.
- Røed, Ragnhild Klingenberg et al. (2023). "Enhancing questioning skills through child avatar chatbot training with feedback". In: *Frontiers in Psychology* 14.
- Shute, Valerie and Brendon Towle (2018). "Adaptive e-learning". In: *Aptitude*. Routledge, pp. 105–114.
- Shute, Valerie J (2008). "Focus on formative feedback". In: *Review of educational research* 78.1, pp. 153–189.
- Shvets, Oleg, Kristina Murtazin, and Gunnar Piho (2020). "Providing feedback for students in e-learning systems: a literature review, based on IEEE explore digital library". In: *2020 IEEE global engineering education conference (EDUCON)*. IEEE, pp. 284–289.
- Streicher, Alexander and Jan D Smeddinck (2016). "Personalized and adaptive serious games". In: *Entertainment Computing and Serious Games: International GI-Dagstuhl Seminar 15283, Dagstuhl Castle, Germany, July 5-10, 2015, Revised Selected Papers*. Springer, pp. 332–377.
- Susi, Tarja, Mikael Johannesson, and Per Backlund (2007). "Serious games: An overview". In.
- Westera, Wim et al. (2020). "Artificial intelligence moving serious gaming: Presenting reusable game AI components". In: *Education and Information Technologies* 25, pp. 351–380.
- Wieringa, Roel J (2014). *Design science methodology for information systems and software engineering*. Springer.

- Yan, L et al. (2023). "Practical and ethical challenges of large language models in education: A systematic literature review". In: *arXiv preprint arXiv:2303.13379*. URL: <https://arxiv.org/abs/2303.13379>.
- Yang, Kai-Hsiang (2018). "Learning behavior and achievement analysis of a digital game-based learning approach integrating mastery learning theory and different feedback models". In: *Learning Analytics*. Routledge, pp. 93–106.
- Yannakakis, Georgios N and Julian Togelius (2018). *Artificial intelligence and games*. Vol. 2. Springer.
- Yi, Siqi et al. (2017). "The role of feedback in game-based learning: A review of the literature". In.
- Yin, Shukang et al. (2023). "Woodpecker: Hallucination correction for multimodal large language models". In: *arXiv preprint arXiv:2310.16045*.
- Zeng, Jialing, Sophie Parks, and Junjie Shang (2020). "To learn scientifically, effectively, and enjoyably: A review of educational games". In: *Human Behavior and Emerging Technologies 2.2*, pp. 186–195.