HEILBRONN UNIVERSITY
OF APPLIED SCIENCES

Master Thesis

Title

Assessing Classroom Management through AI: Comparing LLM-Based and Human Evaluations in Virtual Teaching Simulations

Version 1.0
02. February 2026

Supervision
Prof. Dr.-Ing. Gerrit Meixner
Mario Schwarz, M.Sc.

Marsal
Martin
209390
Software Engineering
Heilbronn University

I. Abstract

The evaluation of classroom management skills is a time-consuming bottleneck in teacher education. This master thesis addresses this challenge by implementing an automated feedback system driven by the GPT-5.1 Large Language Model (LLM) within *VR-Academy*, a virtual reality (VR) application designed for training pedagogical skills in conflict situations through interaction with virtual students.

To evaluate the system, a comparative user study was conducted with pre-service teachers ($n = 7$) and non-teachers ($n = 7$). Participants completed a teaching simulation and subsequently evaluated feedback generated by the Artificial Intelligence (AI) against feedback provided by a human supervisor.

Statistical analysis revealed a significant difference favoring human feedback only for Perceived Usefulness ($p < .05$). Although other dimensions did not reach statistical significance, large effect sizes (r_{rb} from $.61$ to $.81$) were observed. Descriptive trends indicate that the human feedback was perceived as more helpful, motivating, and closer to the optimal length, while the AI feedback was rated as more detailed and its vocal delivery was perceived as pleasant.

Importantly, although the human feedback generally achieved higher absolute ratings, the scores for both feedback types remained high, suggesting that the LLM-based system has reached a level of maturity sufficient for educational use. Notably, no significant differences were found between the ratings of pre-service teachers and non-teachers, despite some medium-to-large effect sizes ($r_{\text{rb}} > .30$), indicating that the tool is accessible to varying levels of expertise.

A decisive finding was the unanimous preference (100%) for a combination of both feedback types. This implies a future paradigm where AI facilitates high-frequency, low-stakes practice, allowing human mentors to focus on complex, motivational coaching. Furthermore, the modular architecture developed in this thesis serves as a scalable blueprint for deploying automated feedback across future scenarios in the VR-Academy.

II. Contents

Title **1**
I. Abstract **2**
II. Contents **3**
III. List of Figures **6**
IV. List of Tables **6**
V. List of Diagrams **7**
VI. List of Code **7**
VII. List of Callouts **7**
1. Introduction **1**
1.1. Problem Statement 1
1.2. Methodology 1
1.3. Research Questions 2
1.4. Hypotheses 2
2. Related Work **3**
2.1. Virtual Reality in Teacher Education 3
2.2. Cultivation of Teachers' Reflection Skills in VR 4
2.3. The Cognitive Affective Model of Immersive Learning (CAMIL) 4
2.4. Extended Reality (XR) in Teacher Education 5
2.5. Large Language Models & Artificial Intelligence for Behavior Assessment . . 6
2.6. Theoretical Foundations of VR-Based Reading 7
2.7. Analysis of Top-Performing Models 8
2.8. Techniques for Prompt Engineering 10
2.8.1. Prompt Engineering Strategies 10
2.8.2. Few-Shot Prompting 11
2.9. Perceived Usefulness 11
3. Implementation **13**
3.1. Introduction to TURTLE 13
3.2. VR-Academy 13
3.3. Integrating AI into a Theoretical Framework 15
3.4. Flags 16
3.5. Conversation Histories 17
3.6. Constructing the Prompt 18
3.6.1. Introductory Instructions 18
3.6.2. Broadcast Session 19
3.6.3. Individual Chat Sessions 20
3.6.4. Flags 23
3.6.5. Criteria 23
3.6.6. Final Instructions 24
3.7. Cloning the Prompt Asset 24
3.8. Sending the Prompt 25

3.9. Retry Logic of API Calls ... 27
3.10. Communicating AI Feedback to Users ... 28
3.11. Text-to-Speech ... 31
3.12. Operational Overview of the Popup Manager ... 32
3.13. Prompt Engineering ... 34
3.13.1. The Lack of Didactic Feedback ... 34
3.13.2. The Necessity for Descriptive Feedback ... 34
3.13.3. An Improved Prompt Version ... 36
3.13.4. The Final Prompt Version ... 37

4. Study Design ... 38
4.1. General ... 38
4.2. Data Collection ... 38
4.3. Study Procedure ... 41

5. Results ... 44
5.1. Participants ... 44
5.2. AI and Human Ratings ... 46
5.2.1. Quantitative Results ... 46
5.2.2. Perceived Usefulness ... 47
5.2.3. Qualitative Results ... 47
5.3. AI Voice ... 49
5.3.1. Quantitative Results ... 50
5.3.2. Qualitative Results ... 50
5.4. Final Survey ... 50
5.4.1. Comfort ... 51
5.4.2. Helpfulness, Objectivity, Motivation, and Detail ... 51
5.4.3. Future Preference ... 53
5.5. Statistical Analysis ... 54
5.5.1. Mann-Whitney U Test ... 54
5.5.2. Wilcoxon Signed-Rank Test ... 55

6. Discussion ... 57
6.1. AI and Human Ratings ... 57
6.2. AI Voice ... 58
6.3. Comfort ... 59
6.4. Helpfulness, Objectivity, Motivation, and Detail ... 60
6.5. Future Preference ... 61
6.6. Limitations ... 61

7. Conclusion ... 63
7.1. Summary of Findings ... 63
7.2. Implications ... 63

8. Outlook ... 65
8.1. Technical Refinements ... 65
8.2. Methodological Refinements ... 65
8.3. Scalability and Architectural Paradigm ... 66

Bibliography ... 67

Glossary ... 69

Template Information **70**
Affidavit of Martin Marsal **71**
Appendix **1**
A. Flag List **1**
B. First Prompt **3**
C. Second Prompt **6**
D. Third Prompt **10**
E. Fourth and Final Prompt **15**

III. List of Figures

Figure 1 Overview of the CAMIL. Reprinted from Makransky et al. [1], under CC BY 4.0 license [2]; no changes were made. 5

Figure 2 Visual representation of the five virtual avatars. Reprinted from Nygren et al. [3], under CC BY 4.0 license [2]; no changes were made. 6

Figure 3 Comparison of frontier LLMs ranked by the Artificial Analysis Intelligence Index [4]. 8

Figure 4 Comparison of LLM inference speeds based on output tokens per second [4]. 9

Figure 5 Cost efficiency comparison of LLMs based on price per million tokens [4]. 10

Figure 6 Overview of the VR-Academy classroom layout and student seating arrangement. 14

Figure 7 Representation of a disruptive classroom incident featuring antisemitic writings (German) on the virtual blackboard. 14

Figure 8 Visual representation of the user interface during a direct interaction with student Leon. 15

Figure 9 User interface showing feedback based on the first criterion 'Target Group Focus'. 30

Figure 10 Participants' experience with Virtual Reality. 45

Figure 11 Participants' experience with simulations in educational contexts. 45

Figure 12 Participants' experience with video games. 46

Figure 13 Participant comfort levels regarding AI evaluation. 51

Figure 14 Pre-service teacher preferences for AI vs. Human feedback stratified by aspect. 52

Figure 15 Non-teacher preferences for AI vs. Human feedback stratified by aspect. 52

IV. List of Tables

Table 1 Participant demographics. 44

Table 2 Comparison of AI and human feedback ratings across both groups (Mean [Median]) 47

Table 3 Participant ratings of the AI-generated feedback voice. 50

Table 4 Statistical differences in feedback ratings between pre-service teachers and non-teachers (Mann-Whitney U). 55

Table 5 Wilcoxon Signed-Rank comparison of AI vs. human feedback ratings. 56

Table 6 A complete list of all unlockable flags in the first scene. 1

V. List of Diagrams

Diagram 1 Retry logic with exponential backoff. 27
Diagram 2 Orchestrating input, feedback display, and TTS via FeedbackPopupTextManager.cs. 33
Diagram 3 Overview of the study procedure. 42

VI. List of Code

Code Snippet 1 FeedbackPromptData.cs defining the structure of the prompt. 16
Code Snippet 2 Flag.cs defining the structure of a flag. 17
Code Snippet 3 One-on-one chat sessions being stored (previous implementation). 18
Code Snippet 4 Broadcast chat sessions being stored (previous implementation). . 18
Code Snippet 5 Initialization of the first prompt component. 19
Code Snippet 6 Algorithm for integrating broadcast session data into the prompt. 20
Code Snippet 7 Algorithm for integrating individual session data into the prompt. .21
Code Snippet 8 Algorithm for integrating the flags. 23
Code Snippet 9 Integrating the criteria into the prompt. 24
Code Snippet 10 Integrating the final instructions into the prompt. 24
Code Snippet 11 Cloning the TeacherFeedbackPrompt.asset 25
Code Snippet 12 Prompt sanitization in the methods SendToOpenAI and SendToHHN. 26
Code Snippet 13 FeedbackEntry class defining the structure of a list entry. 26
Code Snippet 14 InitAsync method in FeedbackPopupTextManager.cs. 28
Code Snippet 15 ShowCurrent method in FeedbackPopupTextManager.cs. 29

VII. List of Callouts

Callout 1 Few-shot prompting paradigm featuring English-to-French translation exemplars. Adapted from Brown et al. [5]. 11
Callout 2 Incorrect format of the broadcast message of an NPC. 19
Callout 3 Correct format of the broadcast message of an NPC. 19

Callout 4 Initial storage of a user message in ChatSessionManager.cs. 22
Callout 5 Example output of the chat messages in the prompt. 22
Callout 6 Example output of the dLeviTalkedto flag. 23
Callout 7 Transcript of the feedback for the criterion 'Target Group Focus'. 30
Callout 8 Previous voice instructions for the feedback texts. 31
Callout 9 Final voice instructions for the feedback texts. 32
Callout 10 One-shot-prompting: Usage of a demonstration in the prompt. 36
Callout 11 Pre-study survey focusing on demographics and prior experience. 39
Callout 12 Survey evaluating the AI feedback and human feedback. 40
Callout 13 Survey evaluating the AI-generated voice. 41
Callout 14 Final survey comparing AI vs. human feedback. 41

1. Introduction

1.1. Problem Statement

The evaluation of prospective teachers is a critical process, typically conducted by third parties such as experienced educators or school administrators. A central criterion in this assessment is classroom management, defined here as the way a teacher interacts with students, how they proactively prevent disruptions, and whether they maintain a friendly, yet consistent presence.

Traditionally, evaluating these complex skills has relied on established frameworks like the Flanders Interaction Analysis System (FIAS) [6]. However, this manual approach is time-consuming and resource-intensive.

In parallel, Virtual Reality (VR) has emerged as a powerful tool for teacher training in the last decade, offering immersive environments where educators can practice conflict resolution in a safe, controlled setting [7]. However, scalable training in VR is often limited by the need for human supervision to provide feedback.

With the rapid development of Large Language Models (LLMs) in recent years, these tools have accelerated work processes in many areas and reduced people's workloads, including the evaluation of tutors in training [8]. This thesis investigates the synergy of these technologies: can an LLM-driven system within a VR environment provide feedback on complex teaching behaviors, specifically in discrimination conflicts that is comparable to human evaluation? By combining the immersion of VR with the analytical capabilities of LLMs, the implementation and the study in this thesis explore a new paradigm for automated, scalable teacher assessment.

1.2. Methodology

Initially, a comprehensive literature review was performed across different research domains, specifically teacher evaluation, VR, and LLMs.

Subsequently, a feedback logic was integrated into a pre-existing Unity VR project. This project features a classroom-based VR scenario where the player is required to interact with virtual students and mitigate discriminatory conflicts. The technical implementation necessitated the extraction of data from a dialogue system and its transmission to an LLM to generate feedback. This process concluded with an "island style" character presenting the final results.

Following the implementation phase, a user study was conducted involving two distinct groups: pre-service teachers and non-teachers. Both groups completed the VR scenario

and were evaluated by the Artificial Intelligence (AI) -driven feedback logic. Concurrently, participants were assessed by a human supervisor to facilitate a direct comparison between the two methods. This design allowed participants to assess the relative value of AI -generated feedback versus human feedback.

1.3. Research Questions

The research questions were the following:

1. What design decisions regarding prompting and structure are made during an iterative development process with stakeholders to optimize the perceived quality of the LLM feedback?
2. To what extent can LLMs provide high-quality, formative feedback on teachers' reactions to classroom discrimination?
3. How does LLM-based assessment compare to traditional assessment methods?
4. What are the qualitative differences in the feedback provided by LLMs compared to human assessors?
5. Do participants prefer to be assessed by AI or humans?

1.4. Hypotheses

The hypotheses were the following:

H1: There is no significant difference in Perceived Usefulness between the AI feedback and the human feedback.

H2: The perceived length of human feedback will be rated as significantly more appropriate (closer to the optimal score of "3") than the AI feedback.

H3: Participants will rate the AI voice as unpleasant (mean rating below 3).

H4: Participants will rate the AI feedback as significantly more detailed than the human feedback.

H5: A combination of both AI and human feedback will be preferred.

2. Related Work

To establish an overview of related work and the current state of the art, a systematic literature review is presented in the following chapter. The majority of sources were identified using Google Scholar. While publication dates and literature types were generally not restricted, a specific temporal constraint was applied to references concerning LLMs. For these, priority was given to sources published within the last three years to ensure relevance in this rapidly evolving field.

2.1. Virtual Reality in Teacher Education

Examining the expansion of VR in recent years is essential to determine its efficacy in enhancing teacher education. A publication by Huang et al. [7] provides a systematic review of the state-of-the-art literature in this domain, utilizing specific inclusion criteria: studies had to involve VR for teacher education, be peer-reviewed (qualitative or quantitative), and be published in English between 2010 and 2020.

Out of 48 identified studies, results indicate a significant surge in VR technology between 2014 and 2016, with over half of the research published between 2019 and 2020. Pre-service teachers constituted 80% of the participants across 46 studies, signaling a demand for VR training within this group, a demographic likewise prioritized in this thesis.

Regarding technological immersion, 41.3% of the studies were classified as fully-immersive, 34.8% as semi-immersive, and 23.9% as non-immersive. The different classifications are defined as follows:

- Fully-immersive: Completely covers the field of view and blocks out visual surroundings for the user via a Head-Mounted Display (HMD) .
- Semi-immersive: Uses wall-mounted displays and motion tracking devices.
- Non-immersive: Uses traditional devices such as desktop monitor, keyboard, mouse, etc. to enable interaction with the VR environment.

Based on these metrics, the VR-Academy project and its associated study align with the fully-immersive category. Application themes ranged from lesson planning to managing student disruptions. Notably, only four studies addressed high-risk interventions involving diversity, homelessness, or substance use. This scarcity highlights that conflict-resolution scenarios remain underrepresented, underscoring the significance of this research.

The studies categorized outcomes into factual/conceptual knowledge (3 studies), procedural knowledge (38 studies), and uncategorized outcomes (10 studies). Procedural knowledge, which is defined as knowing “how and when” to act, was central to eight

studies focused on classroom management. Despite this focus, the integration of AI within classroom management research remains a critical gap.

While most studies reported positive effectiveness, primarily via self-reports and observations, it is important to note that this review concludes in 2020. Hence, to extend these findings, Han et al. [9] conducted a systematic literature review on the use of VR in teacher education from 2014 to 2024. Their analysis included 52 empirical studies, 22 of which employed experimental or quasi-experimental designs. The inclusion criteria required that participants be teachers and that VR serve as a primary training tool.

The results indicate a surge in publications after 2020, likely driven by advancements in VR technology and the shift toward digital learning necessitated by the COVID-19 pandemic. Consistent with Huang et al. [7], 83% of participants were pre-service teachers, with the remainder consisting of in-service teachers or mixed cohorts. The most frequent learning objective identified was classroom management. Furthermore, the meta-analysis revealed that VR achieved moderate overall effectiveness in teacher education.

2.2. Cultivation of Teachers' Reflection Skills in VR

Within the review by Huang et al. [7], the study by Stavroulia et al. [10] is identified as a significant contribution regarding VR systems designed for addressing racial diversity and verbal bullying. This research evaluates the reflection skills between in-service teachers in traditional classroom settings and those using a VR environment. The thematic focus was derived from a literature review indicating that critical incidents, particularly racism and bullying, represent the most challenging scenarios for educators.

The experimental design involved the introduction of a new foreign student who encounters verbal bullying. Participants could observe the interaction from two perspectives: the victimized student and the teacher. Given that most participants had prior teaching experience, a questionnaire was employed to quantify their reflection.

The findings demonstrate that while reflection is essential for all participant groups, the VR system facilitated a perspective, allowing teachers to experience the scenario as the student, which resulted in a higher degree of reflection. This suggests that VR offers pedagogical advantages over traditional classroom environments by enabling multi-perspective immersion.

2.3. The Cognitive Affective Model of Immersive Learning (CAMIL)

In 2021, Makransky et al. [1] introduced the Cognitive Affective Model of Immersive Learning (CAMIL) , a framework designed to describe the learning process within immersive VR environments. As this model serves as a theoretical foundation for a later article [11], its core components are detailed here. A conceptual overview of the model is provided in Figure 1.

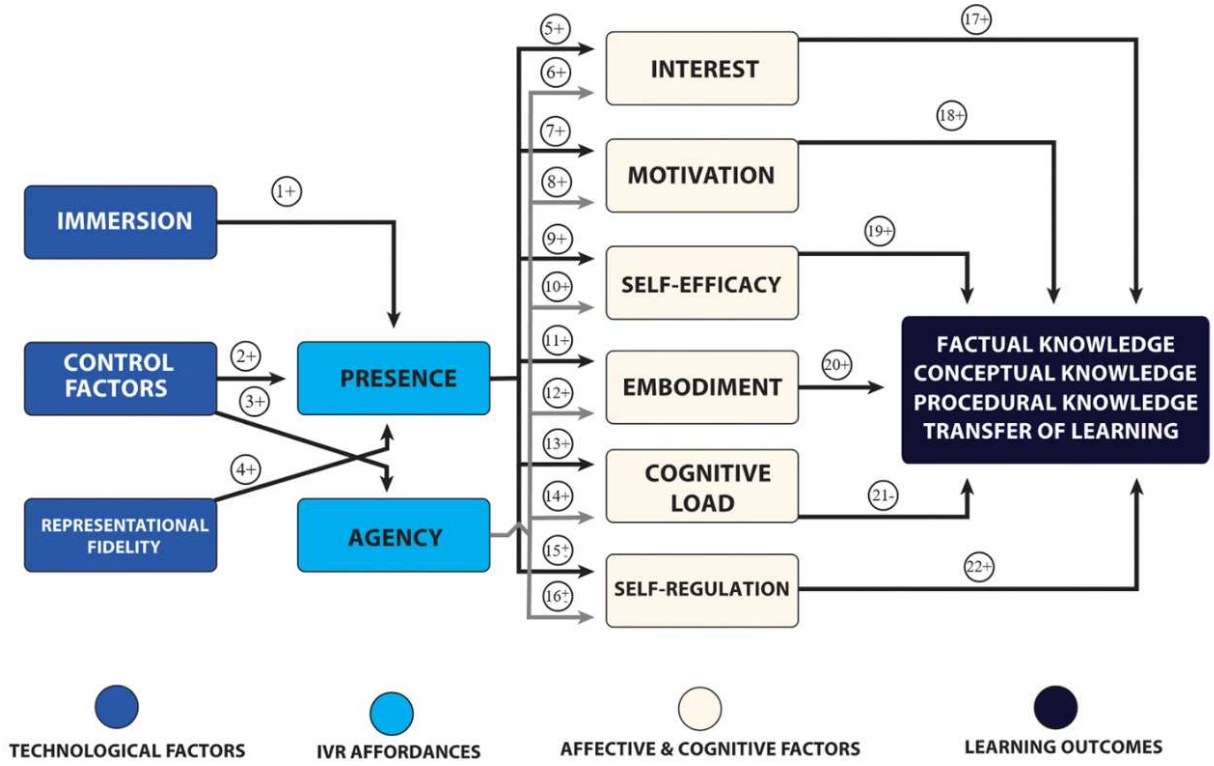

Figure 1: Overview of the CAMIL. Reprinted from Makransky et al. [1], under CC BY 4.0 license [2]; no changes were made.

The framework posits that the primary psychological affordances of immersive learning, namely presence and agency, are influenced by technological determinants, specifically immersion, control factors, and representational fidelity. These affordances subsequently influence several affective and cognitive mediators: interest, motivation, self-efficacy, embodiment, cognitive load, and self-regulation. Ultimately, the model identifies four critical learning outcomes: factual, conceptual, and procedural knowledge, alongside transfer of learning.

2.4. Extended Reality (XR) in Teacher Education

To further extend the comparison with the review by Huang et al. [7] and incorporate literature beyond 2020, Wang et al. [11] conducted a systematic review focusing on teacher education within VR , Augmented Reality (AR), and Mixed Reality (MR), collectively referred to as Extended Reality (XR) .

The authors queried six databases in November 2022: Web of Science, Scopus, IEEE Xplore, ERIC, ScienceDirect, and the ACM Digital Library. From an initial yield of 1490 articles, a systematic filtering process resulted in 52 relevant studies. To categorize the training objectives within these studies, the authors applied the CAMIL framework [1].

Among these, 32 studies evaluated procedural knowledge, with 14 utilizing XR technologies for classroom management training. Specifically, four focused on “professional noticing”. Other targeted pedagogical skills included parent-teacher communication, identifying student needs, and test anxiety awareness.

Twelve additional studies investigated psychological dimensions, seven of which assessed teacher self-efficacy following XR interventions. Other psychological variables included stress reactions, affective states, and emotional affordances. The remaining research focused on conceptual knowledge, covering topics such as discrete trial training, proportional reasoning, and cardiac anatomy.

2.5. Large Language Models & Artificial Intelligence for Behavior Assessment

A study closely aligned with this research was published by Nygren et al. in 2025 [3]. The authors conducted mixed-reality simulations involving pre-service teachers who engaged with five student avatars to facilitate discussions on sensitive and controversial topics. The specific classroom configuration for this environment is illustrated in Figure 2.

Figure 2: Visual representation of the five virtual avatars. Reprinted from Nygren et al. [3], under CC BY 4.0 license [2]; no changes were made.

Each student avatar possesses a distinct personality, mirroring the design of the VR-Academy project. Following the dialogue, both AI and human evaluators analyzed the transcripts to provide feedback, utilizing Shulman's framework to classify and critique participant actions. This feedback was structured via coding schemes, including categories such as content knowledge and pedagogical content knowledge. The study employed three distinct LLMs: ChatGPT-4, ChatGPT-4o, and Claude 3.5 Sonnet.

The findings indicate that human experts generally exhibited higher rating consistency than the AI models. This discrepancy may stem from the fact that AI prioritizes different conversational dimensions than human evaluators. While humans effectively identified missed pedagogical opportunities, the AI struggled to categorize knowledge-based aspects according to Shulman's framework. Additionally, the AI demonstrated a pronounced positivity bias. Specifically, 100% of ChatGPT's evaluations were positive, whereas human feedback remained more mixed.

In a separate study by Thomas et al. [8], tutors were prompted to respond to common instructional scenarios. Notably, this investigation utilized text-based applications rather

than VR environments. One scenario involved a student who frequently withdraws after initial failure but eventually solves a math problem. The tutor is then tasked with providing encouragement to foster persistence. Subsequently, generative AI was used to provide feedback and assessment of the tutor's intervention. The AI offered constructive suggestions and proposed optimized alternative responses. For the assessment phase, 50 tutor-student dialogues concerning math errors were analyzed. The LLM was able to assess the criteria related to successfully responding to an error of the student.

While the AI yielded promising results, the authors highlighted practical and ethical considerations. A "human-in-the-loop" approach remains essential to ensure the quality and accuracy of AI outputs. Furthermore, the development of these systems necessitates strategic AI methodologies and should incorporate direct input from the primary stakeholders, specifically the tutors.

Gao et al. [12] demonstrated how tracking data from a HMD and its controllers can assess teacher expertise via machine learning models. Their user study required both experienced and pre-service teachers to deliver presentations in a VR classroom while managing disruptive student behavior. The researchers evaluated three classification models: Support Vector Machine (SVM), Random Forest, and LightGBM. Among these, the Random Forest model achieved the highest accuracy, yielding a ROC-AUC score of 0.768. Although this study prioritizes non-verbal gestures over verbal communication and does not employ LLMs, it illustrates an alternative methodology for utilizing AI to evaluate pedagogical behavior within virtual environments.

2.6. Theoretical Foundations of VR-Based Reading

Given that the feedback system developed for this thesis necessitates extensive text consumption, it is critical to investigate how various User Interface (UI) designs influence the comfort and efficacy of reading long-form content within VR .

In a within-participants study from 2023, Gabel et al. [13] evaluated four different UI designs for immersive reading:

1. Continuous text with a scrollbar.
2. Continuous text with a scrollbar and up/down buttons.
3. Discrete texts with vertical pagination and arrow buttons.
4. Discrete texts with horizontal pagination and arrow buttons.

Continuous variants presented text on a single scrollable pane, whereas discrete designs divided content into navigable pages. Participants interacted with four distinct texts, with both content and UI order randomized. The evaluation utilized standardized instruments, specifically the Simulator Sickness Questionnaire (SSQ) and the User Experience Questionnaire (UEQ), alongside qualitative feedback and recall metrics (Information and Spatial Recall) to assess row-level accuracy.

The findings revealed no statistically significant differences in objective reading performance. However, notable variations emerged in user experience. Both discrete pagination variants outperformed continuous designs in attractiveness and several

UEQ dimensions. Conversely, the continuous scrollbar variant received the lowest overall rankings. Consequently, the authors recommend pagination-based patterns over traditional scrollbar interfaces for VR applications.

2.7. Analysis of Top-Performing Models

When selecting LLMs for specific use cases, it is essential to consult contemporary performance leaderboards. Due to the rapid pace of model iteration, acquiring the most recent evaluative data is critical, particularly when integrating these models into production-level applications.

Figure 3 illustrates a leaderboard from Artificial Analysis [4], a platform that provides frequent analyses and updates based on state-of-the-art model performance. As of November 2025, the “Intelligence” leaderboard is led by Gemini 3 Pro Preview, GPT-5.1 (high), Kimi K2 Thinking, Grok 4, and Claude 4.5 Sonnet, which hold the top five positions with Intelligence scores of 73, 70, 67, 65, and 63, respectively. The Intelligence Index serves as a benchmark, evaluating models across multiple technical dimensions, including reasoning, knowledge acquisition, mathematics, and programming.

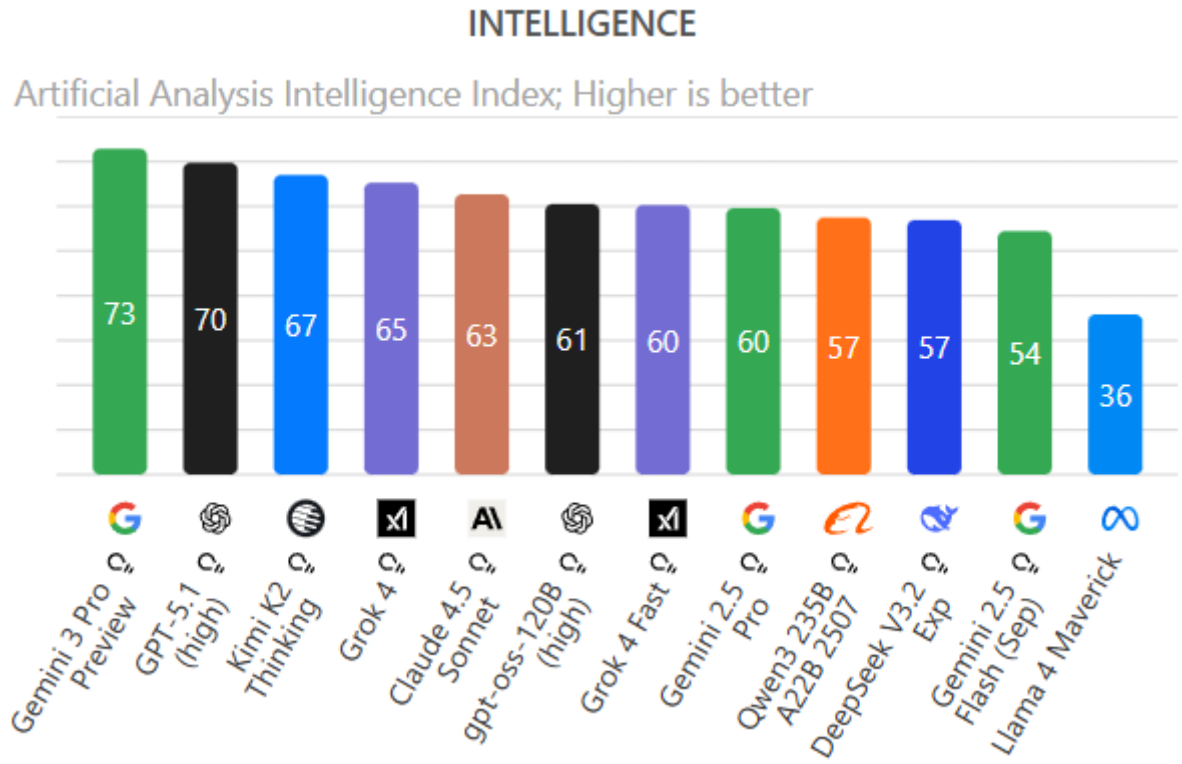

Figure 3: Comparison of frontier LLMs ranked by the Artificial Analysis Intelligence Index [4].

Figure 4 illustrates the highest-performing models in terms of output throughput, measured in tokens per second. This metric quantifies the rate at which a model generates discrete units of text where a token may represent a single word or a sub-word character during response generation. According to the data, the leading five models are gpt-oss-120B (high), Grok 4 Fast, GPT-5.1 (high), Gemini 2.5 Flash (Sep), and Llama

4 Maverick, exhibiting generation speeds of 339, 179, 152, 149, and 128 tokens per second, respectively.

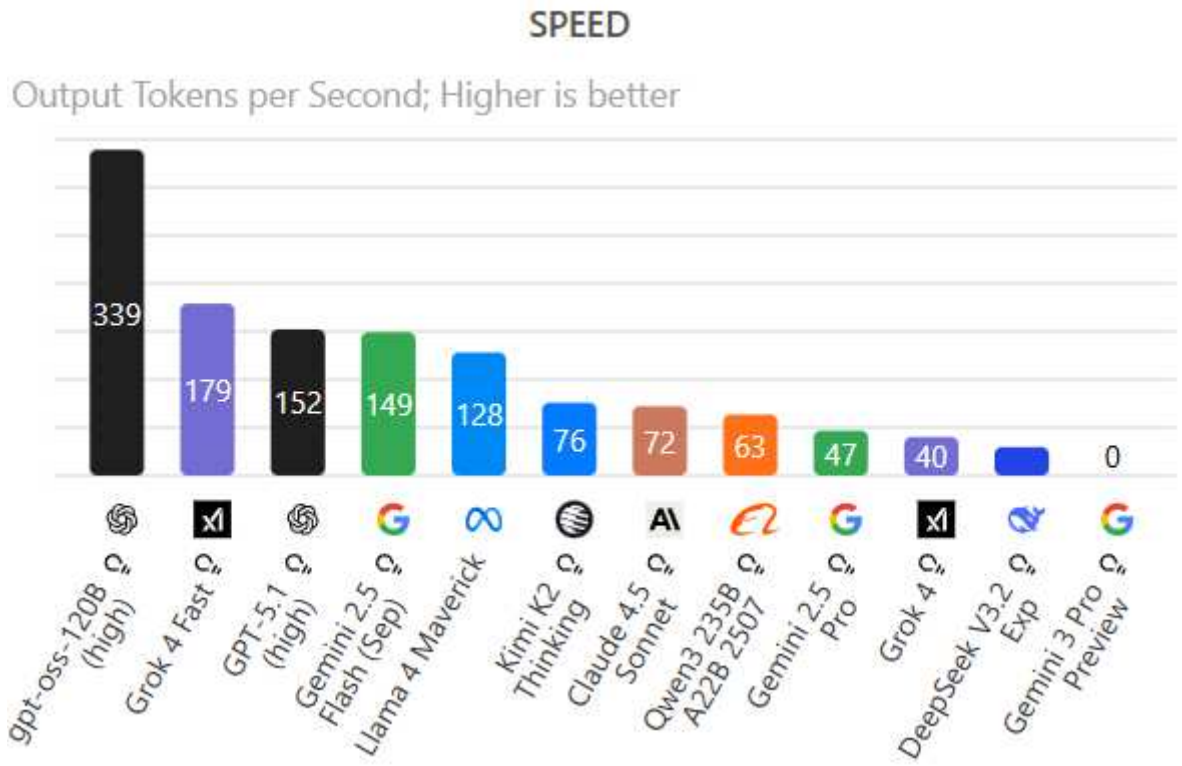

Figure 4: Comparison of LLM inference speeds based on output tokens per second [4].

The selection of an appropriate model is heavily contingent upon the specific requirements of the application's use case. While high-speed generation is critical for interactive, real-time interfaces to ensure a seamless user experience, other scenarios may prioritize reasoning depth or output quality, where a higher latency is acceptable.

Lastly, Figure 5 displays the most cost-effective models, quantified by USD per one million tokens. The five leading models in this category are gpt-oss-120B (high), Grok 4 Fast, DeepSeek V3.2 Exp, Llama 4 Maverick, and Gemini 2.5 Flash (Sep), with respective costs of 0.3, 0.3, 0.3, 0.4, and 0.8 USD.

As previously stated, model selection must align with the specific requirements of the application. In scenarios where high cognitive performance is necessary and budgetary constraints are secondary, the Intelligence Index serves as a primary reference. Conversely, if an application necessitates a balance of low latency and cost-efficiency, an optimal trade-off between the speed and pricing leaderboards must be established.

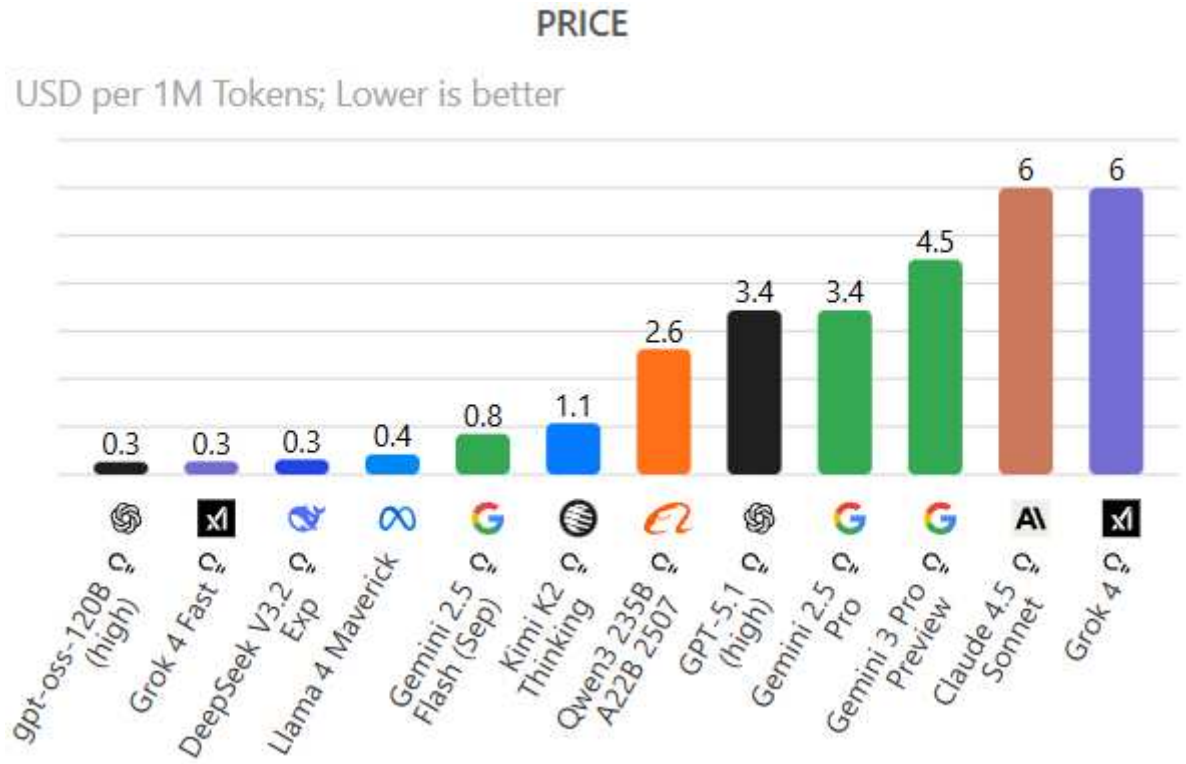

Figure 5: Cost efficiency comparison of LLMs based on price per million tokens [4].

2.8. Techniques for Prompt Engineering

Developing effective prompts for specialized use cases presents a significant challenge. Due to the nondeterminism of LLMs, achieving a consistent output often requires an iterative process of refinement. The subsequent sections demonstrate several strategies and techniques designed to optimize prompt performance. These methodologies represent a critical framework for the technical implementation of the feedback system discussed in later chapters.

2.8.1. Prompt Engineering Strategies

In 2023, Indran et al. [14] published a guide outlining twelve strategies for utilizing ChatGPT to generate medical exam questions efficiently. While the primary context, namely medical assessment, differs from generating feedback for virtual teaching simulations, several principles are universally applicable to prompt engineering.

The following strategies are particularly relevant to the technical implementation of this thesis:

1. Utilizing the most recent model is recommended to enhance output quality and accuracy, though potential subscription costs must be considered.
2. Instructions should be organized as numbered lists rather than dense, continuous paragraphs to improve model adherence.

3. Iterative refinement of terminology is essential, as specific word choices can significantly alter the model's generated response.
4. The technical limitations of LLMs must be integrated into the prompt design to manage expectations and output validity.

2.8.2. Few-Shot Prompting

In 2020, researchers at OpenAI, led by Brown et al. [5], introduced a technique known as few-shot prompting. This methodology involves providing the model with "demonstrations", which are specific examples of the desired output, to significantly enhance the quality and consistency of the generated results.

One-shot prompting follows a similar logic but utilizes only a single demonstration. Conversely, zero-shot prompting provides no examples, relying entirely on a natural language description of the task. Although few-shot prompting often yields superior accuracy, one-shot or zero-shot approaches may be more appropriate or sufficient depending on the complexity of the specific context.

Callout 1 illustrates a few-shot prompting application as described by Brown et al. [5], where several demonstrations follow the initial task instruction.

Few-Shot Prompting Example

Translate English to French:

sea otter => loutre de mer

peppermint => menthe poivrée

plush girafe => girafe peluche

cheese =>

Callout 1: Few-shot prompting paradigm featuring English-to-French translation exemplars. Adapted from Brown et al. [5].

Notably, unlike fine-tuning, which involves updating model weights, these prompting techniques rely exclusively on the provided context without altering the underlying model parameters.

2.9. Perceived Usefulness

To ensure the scientific validity of the metrics and questions used in the subsequent user study, established frameworks must be utilized. In 1989, Davis [15] introduced the Technology Acceptance Model (TAM) , which includes validated scales for two primary variables: Perceived Usefulness (PU) and Perceived Ease of Use (PEOU). These constructs were empirically validated through two studies involving 152 participants. For the scope

of this thesis, assessing the Perceived Usefulness of the proposed feedback system is particularly relevant.

The original scale items for the Perceived Usefulness variable include:

1. My job would be difficult to perform without electronic mail.
2. Using electronic mail gives me greater control over my work.
3. Using electronic mail improves my job performance.
4. The electronic mail system addresses my job-related needs.
5. Using electronic mail saves me time.
6. Electronic mail enables me to accomplish tasks more quickly.
7. Electronic mail supports critical aspects of my job.
8. Using electronic mail allows me to accomplish more work than would otherwise be possible.
9. Using electronic mail reduces the time I spend on unproductive activities.
10. Using electronic mail enhances my effectiveness on the job.
11. Using electronic mail improves the quality of the work I do.
12. Using electronic mail increases my productivity.
13. Using electronic mail makes it easier to do my job.
14. Overall, I find the electronic mail system useful in my job.

Participants indicate their level of agreement with these statements using a Likert scale. Although the original items specifically reference an electronic mail system, selected items can be systematically adapted and refined to align with the context of this research.

3. Implementation

The following chapter demonstrates the technical implementation of the AI-assisted feedback system within the VR-Academy project, which served as the primary tool for the subsequent user study.

3.1. Introduction to TURTLE

TURTLE [16], an acronym for “virTUal RealiTy LErnlandschaft” (virtual reality learning landscape), is an interdisciplinary research initiative led by LMU Munich, FAU Erlangen-Nürnberg, Eberhard Karls University Tübingen, and Heilbronn University of Applied Sciences. Developed as a standalone Unity application for the Meta Quest 3, the project currently investigates immersive pedagogical processes through four different VR environments:

- *Down the rabbit hole:* An exploratory experience designed to raise awareness of social media algorithms. The module addresses how recommendation systems can amplify social polarization, guiding participants through a fictional “rabbit hole” to manage exposure to problematic content.
- *Hoffmann Experience:* An interactive literary environment inspired by the works of German author E.T.A. Hoffmann. Users engage with the narrative as active participants by solving puzzles and interacting with literary characters.
- *VR-Academy:* This module enables users to refine their conflict-resolution skills regarding discrimination in classroom settings. By communicating with AI-controlled students, participants strengthen anti-discrimination and diversity-sensitive competencies before receiving evaluative feedback. This project serves as the central focus of this thesis.
- *DIWA 4.0:* Standing for “Diversity Competence 4.0”, this sub-project provides a virtual space for public service employees and educators to develop anti-racism skills. Participants utilize avatars to reflect on their professional behavior within diverse settings [16].

3.2. VR-Academy

VR-Academy is a project designed to enhance classroom management skills within the context of discrimination and diversity-sensitive scenarios. Currently, the project features a scenario focused on combating anti-Semitism. Users interact with virtual students whose responses are mediated by a dialogue tree system integrated with LLMs. Depending on the appropriateness of the user’s intervention, the dialogue tree either

progresses to the next phase or maintains the current thematic requirement, with the student's response generated accordingly.

Visual documentation of the virtual classroom environment is provided in Figure 6.

Figure 6: Overview of the VR-Academy classroom layout and student seating arrangement.

In this scenario, the user assumes the role of a teacher and is immediately confronted with anti-Semitic writings on the classroom blackboard, as illustrated in Figure 7.

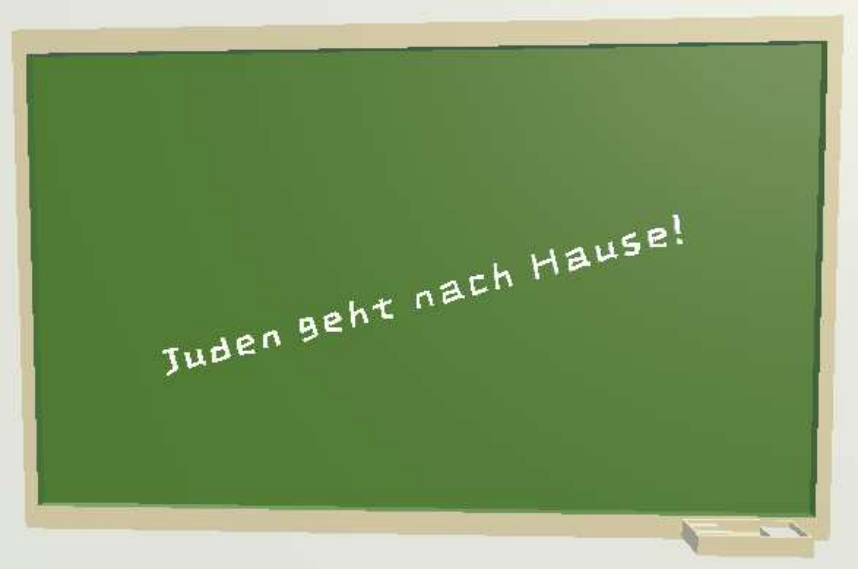

Figure 7: Representation of a disruptive classroom incident featuring antisemitic writings (German) on the virtual blackboard.

The user can engage with three different students to address the conflict:

- *David Levi:* As a Jewish student, David is deeply impacted by the writings and seeks authentic signs of solidarity. He closely monitors the teacher's response and the sincerity of the intervention.
- *Leon Richter:* Although he does not feel directly responsible, Leon is uncertain regarding the appropriate course of action. He attempts to avoid discomfort while simultaneously fearing being perceived as cowardly. Despite having no personal connection to anti-Semitism, he acknowledges its gravity.
- *Mira Steinke:* Recognizing the severity of the issue, Mira aims to facilitate an open classroom discussion. She feels a sense of collective responsibility for the class

atmosphere and desires a shared accountability, even among those not directly involved in the act.

Interaction occurs either through one-on-one dialogue or a “broadcast” mode, where the teacher addresses the entire class and one student at a time responds. The interface for these interactions is illustrated in Figure 8.

Figure 8: Visual representation of the user interface during a direct interaction with student Leon.

The scenario concludes when the user selects the chalk on the teacher’s desk, which triggers the immediate generation of AI -assisted feedback.

The source code for TURTLE and the VR-Academy is hosted in the following repository (subject to authorized access): https://git.it.hs-heilbronn.de/unitylab/turtle/turtle-lerninsel/-/tree/8aa1362dca6745cf8390786a03a61835743edf48.

Given the continuous development of the application, this specific link preserves the project state as it existed for the user study. This commit represents the final iteration of the feedback system prior to the evaluation phase.

3.3. Integrating AI into a Theoretical Framework

The foundational framework for the post-scenario feedback is derived from a theoretical concept developed by the project stakeholders. This concept utilizes “flags” (analogous to digital achievements) to trigger specific feedback. The player is supposed to receive four different feedback texts based on the following criteria/decision levels:

1. Target Group Focus (Zielgruppenfokus): Which students did the player talk to?
2. Problem Naming (Problembenennung): Was the problem explicitly named?
3. Empathy Work (Empathiearbeit): Did the played show empathy to the students?
4. Solution Orientation (Lösungsorientierung): Did the player come up with a solution?

To leverage the capabilities of LLMs, it is essential that the feedback content is not rigidly predefined. Instead, the AI should synthesize the responses dynamically. To bridge

the gap between the original theoretical framework and autonomous generation, a specialized prompt was developed to transmit all relevant application data to the LLM.

3.4. Flags

One critical component of the data transmitted to the LLM is the system of “flags”. A flag functions as an achievement triggered by user interactions. Notably, these can represent both positive pedagogical actions and negative actions, for instance, when a player successfully encourages a student to reflect or proposes a collaborative classroom project. Certain flags are designed to be triggered multiple times.

For the LLM to provide nuanced evaluations, it requires a comprehensive list of both achieved and unachieved flags. This contextual information allows the model to generate constructive feedback by identifying missed opportunities for improvement.

Technically, these flags are integrated into the pre-existing dialogue tree. Previously, the system utilized a HashSet to store triggered flags, which lacked both the frequency of achievements and a record of missing criteria.

To address these limitations, a ScriptableObject titled “FeedbackPromptData.cs” was developed. This script, along with all others related to the feedback system, is located within the directory “VR Lerninsel/Assets/Scripts/VR Academy/Feedback/”.

Code Snippet 1 contains the static configuration for the feedback prompt transmitted to the LLM, including a comprehensive list of all flags. Utilizing the FeedbackPromptData class, an asset can be instantiated via the menu path defined in line 1.

```
1  [CreateAssetMenu(fileName = "FeedbackPromptData", menuName = "LLM/
   Prompt Data", order = 1)]                                          C#
2  public class FeedbackPromptData : ScriptableObject
3  {
4      [TextArea(10, 30)]
5      public string firstPromptPart;  // First part of the prompt until the
       flags section
6
7      public Flag[] flags;            // List of flags with details
8
9      [TextArea(5, 20)]
10     public string criteria;         // Feedback criteria
11
12     [TextArea(10, 30)]
13     public string secondPromptPart; // Second part of the prompt after
       the criteria section
14 }
```

Code Snippet 1: FeedbackPromptData.cs defining the structure of the prompt.

The structure of the Flag class is further detailed in Code Snippet 2.

```C#
[Serializable]
public class Flag
{
    public string flagName;      // e.g. "empathieMitBetroffenen"
    [TextArea] public string description;  // e.g. "Die Lehrperson hat Empathie mit David gezeigt."
    public int triggerCount;     // how many times triggered
    public Criterion criterion; // Dropdown in Inspector
}

[Flags] // Bitmask
public enum Criterion
{
    Zielgruppenfokus = 1 << 0, // Binary: 0001 (Decimal: 1)
    Problembenennung = 1 << 1, // Binary: 0010 (Decimal: 2)
    Empathiearbeit = 1 << 2, // Binary: 0100 (Decimal: 4)
    Lösungsorientierung = 1 << 3  // Binary: 1000 (Decimal: 8)
}
```

Code Snippet 2: Flag.cs defining the structure of a flag.

The Flag class contains the following variables:

- *flagName:* A unique string identifier triggered within the dialogue tree.
- *description:* A short summary of the event associated with the flag. This metadata is essential for the LLM to interpret user actions.
- *triggerCount:* An integer tracking the frequency of the flag's activation.
- *criterion:* A variable mapping the flag to specific evaluative criteria. This enables the LLM to categorize relevant flags for each feedback segment. The corresponding Criterion enum is defined starting at line 11.

The Criterion enum is implemented as a bitmask, where each member corresponds to a unique bit within an integer. This architecture allows for the selection of multiple criteria per flag within the Unity Inspector, as the underlying bitwise values are summed to represent a set of attributes.

A complete catalog of all 25 achievable flags is provided in Table 6 in Appendix A. Since the final prompt is executed in German, the table contents are maintained in the original language to ensure consistency with the AI 's input.

3.5. Conversation Histories

In addition to the flags, the dialogue history represents a critical data source not present in the original theoretical concept. Providing the LLM with the full transcript of all exchanges

between the teacher and students ensures that the generated feedback is grounded in the precise context of the user's interactions.

Prior to the implementation of the feedback system, the ChatSessionManager.cs file, which is located in the directory "VR Lerninsel/Assets/Scripts/VR Academy/", already served as the database for these histories. Consequently, the feedback system's implementation required only the extraction of data from the existing variables within this class.

As previously noted, interactions can occur via one-on-one sessions. These dialogues are maintained within a dictionary structure, as illustrated in Code Snippet 3:

```
1  public Dictionary<NPCProfile, NPCChatSession> sessions = new
   Dictionary<NPCProfile, NPCChatSession>();
```
C#

Code Snippet 3: One-on-one chat sessions being stored (previous implementation).

The keys of this dictionary correspond to the individual student names (David, Leon, Mira), while the values represent the NPCChatSession instances containing the message logs.

As illustrated in Code Snippet 4, interactions conducted before the entire class (broadcasts) are maintained within a separate property of the same NPCChatSession type, mirroring the data structure used for one-on-one sessions.

```
1  public NPCChatSession BroadcastSession { get; private set; }
```
C#

Code Snippet 4: Broadcast chat sessions being stored (previous implementation).

3.6. Constructing the Prompt

Having identified all required data components, the next phase involves the construction of the prompt, as detailed in the following sections. This process is executed within the FeedbackPromptBuilder.cs file. To optimize performance, the StringBuilder class is utilized for prompt construction. This approach ensures efficient string manipulation by avoiding the overhead of creating multiple string objects in memory during the concatenation process.

3.6.1. Introductory Instructions

As previously established in Code Snippet 1, the FeedbackPromptData.cs script defines the structural requirements for the prompt. Based on this class, a ScriptableObject asset titled "TeacherFeedbackPrompt.asset" was instantiated to hold the static data required for generation. The initial segment of the prompt consists of a string containing introductory instructions for the LLM, including an overview of the classroom scenario and concise profiles for each student. The implementation for appending this string is straightforward and is presented in Code Snippet 5.

```
1  StringBuilder sb = new();                                                C#
2
3  var promptData = FeedbackPromptManager.Instance.RuntimeData;
4
5  // First part of the prompt
6  sb.AppendLine(promptData.firstPromptPart);
```

Code Snippet 5: Initialization of the first prompt component.

3.6.2. Broadcast Session

Following the introductory instructions, the dialogue histories are appended to the prompt. Because these data are dynamic and unique to each session, they cannot be stored within the static TeacherFeedbackPrompt.asset and must be extracted from ChatSessionManager.cs at runtime. The algorithm designed to extract the broadcast session is detailed in Code Snippet 6.

A foreach loop, beginning at line 4, iterates through the messages within the broadcast chat session. The primary challenge regarding the data stored in ChatSessionManager.cs is that the raw message format is not optimized for the prompt. For instance, a student message is originally stored as shown in Callout 2.

Broadcast Message (Incorrect Format)

assistant: [Mira Steinke:] Ich finde es toll, dass wir ein Projekt starten. Vielleicht können wir gemeinsam eine starke Botschaft für unsere Klasse setzen.

Callout 2: Incorrect format of the broadcast message of an NPC.

This format includes metadata that is unnecessary for the LLM, such as the “assistant” role prefix and the use of square brackets around student names. To achieve a more legible output, the data is processed into the format shown in Callout 3.

Broadcast Message (Correct Format)

Mira Steinke: Ich finde es toll, dass wir ein Projekt starten. Vielleicht können wir gemeinsam eine starke Botschaft für unsere Klasse setzen.

Callout 3: Correct format of the broadcast message of an NPC.

Consequently, the logic beginning at line 12 of Code Snippet 6 strips the assistant prefix and brackets from NPC messages. If the message originates from the user, as evaluated by the conditional statement in line 6, the format is preserved, as it already aligns with the prompt requirements. Finally, the processed content is appended to the aggregate prompt string at line 28.

```C#
if (chatSessionManager.BroadcastSession != null &&
chatSessionManager.BroadcastSession.chatHistory.Count > 0)
{
    sb.AppendLine("-- Broadcast --");
    foreach (var msg in chatSessionManager.BroadcastSession.chatHistory)
    {
        if (msg.Role == "user")
        {
            sb.AppendLine($"Lehrer: {msg.Content}");
        }
        else
        {
            // Remove "assistant:" prefix if present
            var content = msg.Content.Replace("assistant:", "").Trim();

            // If the text starts with [Name:], extract and reformat
            if (content.StartsWith("[") && content.Contains(":]"))
            {
                int endIndex = content.IndexOf(":]"); // find closing
                bracket with colon
                if (endIndex > 0)
                {
                    // Extract name inside [ ]
                    string name = content[1..endIndex].Trim();
                    // Extract the rest of the text (after ":]")
                    string text = content[(endIndex + 2)..].Trim();
                    content = $"{name}: {text}";
                }
            }
            sb.AppendLine(content);
        }
    }
    sb.AppendLine();
}
```

Code Snippet 6: Algorithm for integrating broadcast session data into the prompt.

3.6.3. Individual Chat Sessions

The extraction of individual NPC chat sessions follows a logic similar to that of the broadcast session. As these dialogues consist of dynamic data unique to each playthrough, they must be retrieved from ChatSessionManager.cs at runtime

rather than being stored within the static TeacherFeedbackPrompt.asset. The specific implementation of this extraction algorithm is detailed in Code Snippet 7.

```C#
if (chatSessionManager.sessions != null)
{
    foreach (var kvp in chatSessionManager.sessions)
    {
        var profile = kvp.Key;
        var session = kvp.Value;

        if (session.chatHistory.Count == 0) continue;

        sb.AppendLine($"-- {profile.npcName} --");
        foreach (var msg in session.chatHistory)
        {
            if (msg.Role == "user")
            {
                // Find last "User Message:" in the content
                var lines = msg.Content.Split('\n');
                foreach (var line in lines)
                {
                    if (line.StartsWith("User Message:"))
                    {
                        // Take only what comes after "User Message:"
                        var cleaned = line.Replace("User Message:",
                        "").Trim();
                        sb.AppendLine($"Lehrer: {cleaned}");
                    }
                }
            }
            else
            {
                sb.AppendLine($"{profile.npcName}: {msg.Content}");
            }
        }
        sb.AppendLine();
    }
}
```

Code Snippet 7: Algorithm for integrating individual session data into the prompt.

This process uses a nested loop structure: an outer foreach loop iterates through the three individual NPC sessions, while an inner loop processes the specific messages within each session.

As with previous data extractions, the raw message format from ChatSessionManager.cs presents an additional challenge. The user's messages are prefixed with the system prompt which is the foundational instructions used during the live simulation. While this inclusion is essential for the dialogue tree to provide the LLM with necessary context during the interaction, it is redundant for the feedback system, which requires only the user's actual spoken input.

An example of this formatting issue, including a truncated version of the system prompt, is illustrated in Callout 4.

Initial User Message

Du bist ein Schüler und bleibst in deiner Rolle. Antworte stets auf Deutsch und relativ kurz (maximal 2 Sätze). Der Nutzer ist deine Lehrperson – sprich ihn oder sie respektvoll mit 'Sie' an, aber verwende keine geschlechtsspezifischen Anreden oder Namen wie 'Herr' oder 'Frau'. Erwähne niemals, dass du eine KI bist. Erwähne niemals [...]

User Message: Hallo David, wie fühlst du dich?

Callout 4: Initial storage of a user message in ChatSessionManager.cs.

For the purpose of generating feedback, only the final line of the user's input, prefixed with "User Message:", is required; all preceding system instructions must be truncated. This parsing logic is implemented between lines 15 and 23 of Code Snippet 7. In contrast, the NPC messages require no further modification and are retrieved in their existing format, as shown in line 29.

An excerpt of the resulting dialogue output for the prompt is provided in Callout 5, which illustrates the processed conversation history with the student David.

Prompt: Chat Messages

=== Gesprächsverläufe ===

-- David Levi --

Lehrer: Hallo David, wie fühlst du dich?

David Levi: Es ist schwer, nicht an die Schmierereien zu denken... Was bedeutet das für uns alle hier?

Callout 5: Example output of the chat messages in the prompt.

3.6.4. Flags

In contrast to the dynamic chat sessions, the list of potential flags is predefined and stored within the TeacherFeedbackPrompt.asset. Throughout the simulation, only the triggerCount field is updated dynamically based on user behavior. As previously discussed in Code Snippet 2, each flag object contains a unique name, a description, one or more evaluative criteria, and an integer of its activation frequency.

The algorithm responsible for extracting and formatting these flags for the LLM is illustrated in Code Snippet 8.

```C#
1   foreach (var flag in promptData.flags)
2   {
3       string statusText = flag.triggerCount > 0
4           ? $"{flag.triggerCount} mal freigeschaltet"
5           : "nicht freigeschaltet";
6
7       sb.AppendLine(
8           $"{flag.flagName}: {flag.description} " +
9           $"Kriterium: {flag.criterion}. " +
10          $"Status: {statusText}"
11      );
12  }
```

Code Snippet 8: Algorithm for integrating the flags.

A foreach loop is used to append each flag to the prompt string. Within this loop, a ternary operator (lines 3–5) is utilized to generate an appropriate description of the triggerCount. Subsequently, the flag's full data is appended to the prompt as detailed in lines 7–10.

An example output for the dLeviTalkedto flag is illustrated in Callout 6.

'dleviTalkedto' Flag Output

dLeviTalkedto: Die Lehrperson hat mit David gesprochen. Kriterium: Zielgruppenfokus. Status: 1 mal freigeschaltet

Callout 6: Example output of the dLeviTalkedto flag.

3.6.5. Criteria

The four evaluative criteria "Target Group Focus", "Problem Naming", "Empathy Work", and "Solution Orientation" are retrieved as a string from the TeacherFeedbackPrompt.asset and appended to the prompt, as detailed in Code Snippet 9.

```
1  sb.AppendLine(promptData.criteria);                                C#
```

Code Snippet 9: Integrating the criteria into the prompt.

3.6.6. Final Instructions

The final segment of the prompt consists of a string that provides final instructions to the LLM. These instructions include critical constraints, such as definition of the output structure and the exclusion of specific terminology. Most importantly, the LLM is directed to format its response as a JavaScript Object Notation (JSON) object comprising five distinct elements: four evaluative feedback texts corresponding to each criterion and a comprehensive summary.

As demonstrated in Code Snippet 10, the implementation of this final concatenation is straightforward.

```
1  sb.AppendLine(promptData.secondPromptPart);                        C#
```

Code Snippet 10: Integrating the final instructions into the prompt.

The initial iteration of the prompt is documented in Appendix B. A later section will demonstrate the iterative refinement process that led from this primary draft to the finalized version used in the study. Regarding the prompt appendices, it should be noted that conversation histories and flag trigger counts are represented by example text, as these components are generated dynamically based on player interactions.

3.7. Cloning the Prompt Asset

Following the implementation of the TeacherFeedbackPrompt.asset, a persistence issue was identified regarding the flag data. Because ScriptableObjects in Unity persist changes made during runtime, the updated triggerCount values were being saved and carried over into the next session. For instance, if a user triggered the “BroadcastCount” flag twice during a playthrough, the counter would initialize at two, rather than zero, at the start of the following session.

To resolve this error, the FeedbackPromptManager.cs script was introduced. The Awake method of this script is illustrated in Code Snippet 11.

In line 9, a runtime instance of the prompt asset is generated via cloning. By using this copy for incrementing flag counters and generating user feedback, the original ScriptableObject remains immutable, ensuring that its initial state, with all counters set to zero, is preserved for subsequent sessions.

Furthermore, any pre-existing instances of the manager are identified and removed via the conditional logic in lines 3 through 6. This implementation follows the Singleton pattern, which ensures that only a single instance of the FeedbackPromptManager class exists at any given time, preventing data conflicts and redundant processing.

```C#
private void Awake()
{
    if (Instance != null && Instance != this)
    {
        Destroy(gameObject);
        return;
    }
    Instance = this;
    RuntimeData = Instantiate(originalData);
}
```

Code Snippet 11: Cloning the TeacherFeedbackPrompt.asset

3.8. Sending the Prompt

Once the user exits the scene and the prompt construction is complete, the data is transmitted to the LLM. This communication logic is implemented within the AcademyFeedbackManager.cs file.

The system supports two distinct service providers, selectable via the Unity Inspector on the FeedbackManager game object:

- *OpenAI:* This provider offers state-of-the-art models, as evidenced by the performance benchmarks in Figure 3. Consequently, OpenAI serves as the primary provider for production and the user study. Specifically, the OpenAI Responses Application Programming Interface (API) is utilized, as it offers significant advantages over the standard Chat Completions API, including optimized performance and reduced latency. For this study, the GPT-5.1 model (gpt-5.1-2025-11-13) was selected. The model operates with default parameters, meaning no reasoning effort is enabled. This ensures the fastest possible response time, typically ranging between 30 and 60 seconds.
- *Heilbronn University (HHN):* Local models are managed via Ollama [17], an open-source framework for hosting LLMs on internal hardware. These are accessible via specific API endpoints or the Open WebUI [18] dashboard. While these localized models may not match the reasoning capabilities of OpenAI's flagship versions, they provide a cost-effective alternative for students at HHN, making them an excellent resource for the development phase.

The execution of these requests occurs within the SendToOpenAI and SendToHHN methods. A shared prerequisite for both providers is prompt sanitization (or JSON string escaping). This step is critical before assembling the request body to prevent formatting errors, as illustrated in Code Snippet 12.

```C#
// Properly escape special JSON characters
string safePrompt = prompt
    .Replace("\\", "\\\\")  // Escape backslashes first
    .Replace("\"", "\\\"")  // Escape double quotes
    .Replace("\n", "\\n")   // Convert newline characters
    .Replace("\r", "\\r")   // Convert carriage returns
    .Replace("\t", "\\t");  // Convert tab characters
```

Code Snippet 12: Prompt sanitization in the methods SendToOpenAI and SendToHHN.

During the assembly of the prompt within FeedbackPromptBuilder.cs, various characters are introduced that are incompatible with the JSON standard. These include backslashes, double quotes, newline characters, carriage returns, and tab characters. To ensure the integrity of the request body, these characters must be escaped, a process that treats them as literal text rather than control characters. This is implemented via the Replace method, which substitutes potentially disruptive characters with their JSON-compliant versions.

To facilitate the structured reception of the five feedback components from the LLM's response, a FeedbackResponse class was defined. This class contains five fields corresponding to each feedback segment. Upon receiving the JSON payload, the data is deserialized into an instance of this class.

Subsequently, these values are mapped to a list of objects, where each entry associates a specific feedback text with its corresponding criterion heading. This structured list serves as the data source for the user interface. The architecture of a list entry is illustrated in Code Snippet 13.

```C#
[System.Serializable]
public class FeedbackEntry
{
    public string heading;
    public string text;

    public FeedbackEntry(string heading, string text)
    {
        this.heading = heading;
        this.text = text;
    }
}
```

Code Snippet 13: FeedbackEntry class defining the structure of a list entry.

3.9. Retry Logic of API Calls

The transmission of the prompt and the subsequent API request represent the most critical phase of the feedback generation process. Any failure during this exchange would terminate the feedback, leaving the user without an evaluation. Potential points of failure include transient network connectivity issues or the unavailability of the API endpoint. To mitigate these risks and ensure system robustness, a retry logic mechanism was implemented within the AcademyFeedbackManager.cs file.

The implemented retry logic is based on a core resilience principle found in large-scale distributed systems like Amazon Web Services (AWS). This principle dictates that an initially unsuccessful operation does not immediately propagate an error. Instead, the operation is automatically retried multiple times to gracefully handle transient failures before escalating to a definitive failure state.

Diagram 1 shows a flowchart that demonstrates how this retry logic works.

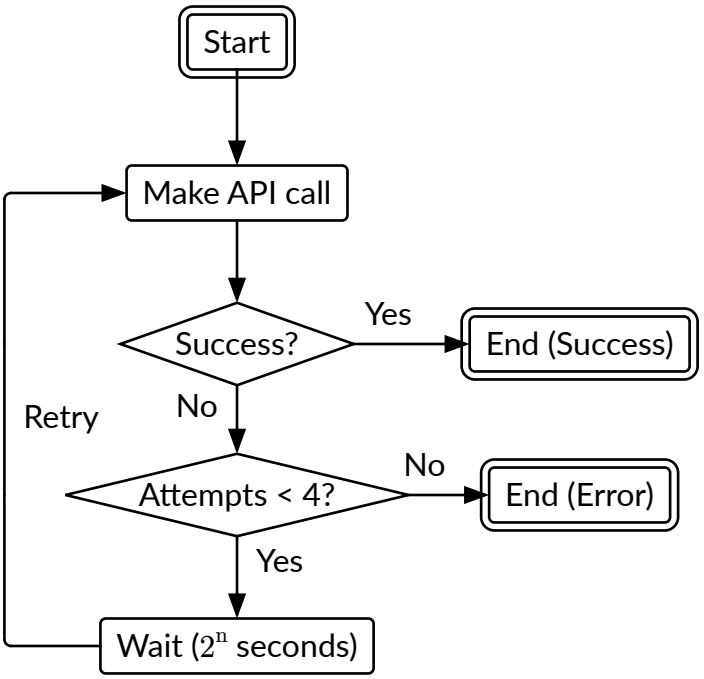

Diagram 1: Retry logic with exponential backoff.

The process begins with an initial API call. If this operation is successful, the process terminates. In the event of a failure, a retry mechanism is initiated that allows for a maximum of four total attempts (the initial call and three retries).

Following a failed attempt, a condition checks if the number of completed attempts is less than four. If this condition is false, the process terminates in an error state. If the condition is true, the system employs an exponential backoff strategy before the next attempt. This strategy introduces a progressively longer delay of 2^n seconds, where n represents the number of the failed attempt (e.g., 2, 4, and 8 seconds). After this delay, the API call is retried, re-initiating the cycle.

In the event of a persistent failure, the user is provided with the option to manually re-initiate the API request. Should the system remain unable to generate feedback

after multiple attempts, a “skip” option is available. This ensures that the user is not permanently obstructed by the feedback interface and can proceed even if the automated evaluation fails.

3.10. Communicating AI Feedback to Users

Once the LLM has returned the feedback texts to the application, they must be presented to the user.

Prior to the implementation of the AI-assisted system, a post-scenario popup was already in place. However, its functionality was limited to displaying basic statistics, such as the total number of interactions and a list of triggered flags. The introduction of generative feedback required a more sophisticated integration to accommodate the feedback texts.

To address this, the FeedbackPopupTextManager.cs file was developed. As the name implies, this file is responsible for managing and dynamically updating the content within the feedback interface. The initialization logic and entry point for this process are detailed in Code Snippet 14.

```C#
1   public async void InitAsync(List<FeedbackEntry> newEntries)
2       {
3           // Import feedback texts
4           entries = newEntries;
5           currentIndex = -1; // -1 is the screen right before playing the
            first feedback text
6
7           // Generate first audio file
8           await feedbackTTS.GenerateSpeech(entries[0].text,
            entries[0].heading);
9
10          feedbackLoadingCircle.SetActive(false);
11          bodyText.text = "Ihr Feedback ist erstellt. Klicken Sie auf
            'Weiter', um es anzuhören.";
12          feedbackNextButton.SetActive(true);
13      }
```

Code Snippet 14: InitAsync method in FeedbackPopupTextManager.cs.

Starting at line 4, the feedback data received from AcademyFeedbackManager.cs (stored as newEntries) is assigned to the local entries variable. The navigation index is initialized at −1. This serves as a transitional state before the first feedback segment (index 0) is presented. During this state, the interface notifies the user that the evaluation has been successfully generated and is ready for review.

The notification text is rendered in line 11, while line 8 triggers the generation of a Text-to-Speech (TTS) voice-over via the OpenAI API. The technical details of this audio integration will be elaborated upon in the following section. When the user interacts

with the “proceed” button, the ShowCurrent method is invoked to transition to the actual feedback content, as illustrated in Code Snippet 15.

```C#
1   private async void ShowCurrent()
2       {
3           headingText.text = entries[currentIndex].heading;
4           bodyText.text = entries[currentIndex].text;
5           headingText.gameObject.SetActive(true);
6
7           await PlayAudio(feedbackTTS.filePath);
8
9           // Pre-generate next audio if this is not the last entry
10          if (currentIndex < entries.Count - 1)
11          {
12              await feedbackTTS.GenerateSpeech(entries[currentIndex +
                1].text, entries[currentIndex + 1].heading);
13              feedbackNextButton.SetActive(true);
14          }
15      }
```

Code Snippet 15: ShowCurrent method in FeedbackPopupTextManager.cs.

The ShowCurrent method updates the interface by mapping the heading text (the criterion) and the body text (the specific feedback) according to the current index, as shown in lines 3 and 4. Following the visual update, line 7 initiates the playback of the associated audio file. To ensure a seamless user experience, the system proactively pre-generates the TTS file for the subsequent feedback segment (lines 10–15), enabling an immediate transition upon the next user interaction.

To enhance immersion, a turtle character asset, sourced from the Unity Asset Store by Nesterov [19], is utilized as a virtual tutor. In combination with the TTS output, the turtle appears to be delivering the feedback directly to the player. The character is positioned next to the popup, oriented toward the player, and utilizes an Animator Controller to loop a continuous swimming animation.

The visual representation of the feedback interface is captured in Figure 9. For improved legibility, a transcript of the feedback, specifically for the “Target Group Focus” (Zielgruppenfokus) criterion, is provided in Callout 7. It should be noted that the content displayed represents the output of the final, optimized prompt; the iterative engineering process required to achieve this quality is detailed in a later section. The user proceeds through the four individual criteria texts sequentially, concluding with the comprehensive summary. As established in Section 2.6 and recommended by Gabel et al. [13], a discrete pagination pattern was implemented in place of scrollable text. This design choice allows the user to navigate the evaluation at their own pace by interacting with the navigation button at the base of the panel.

Figure 9: User interface showing feedback based on the first criterion 'Target Group Focus'.

Feedback for the First Criterion (in German)

Aktion: Sie haben sowohl vor der gesamten Klasse als auch in kurzen Einzelanfragen mit allen drei Schülern gesprochen („Liebe Klasse…“, „David, wie fühlst du dich?“, „Leon, was sagst du…?“, „Mira, hast du eine Idee…?“). Dabei wurde jeder Personengruppe grundsätzlich passend zur Rolle im Geschehen angesprochen: David als direkt Betroffener, Leon als unsicherer Mitschüler, Mira als jemand, der Verantwortung für die Klassengemeinschaft übernehmen möchte.

Positive Implikation: Dies könnte bei David bewirken, dass er wahrnimmt, dass seine Perspektive als Betroffener gesondert beachtet wird und er nicht in der anonymen Gruppe untergeht. Bei Leon könnte es das Signal senden, dass auch seine Sicht zählt und dass er sich nicht hinter der Klasse verstecken muss. Bei Mira könnte diese direkte Ansprache den Eindruck stärken, dass ihr Engagement gewünscht ist und dass sie Ideen einbringen darf. Für die Klasse insgesamt dürfte das Zeichen entstehen, dass Sie verschiedene Rollen im Geschehen unterscheiden.

Kehrseite/Risiko: Gleichzeitig zeigt sich, dass bestimmte Vertiefungen ausgeblieben sind: Der Fokus auf David als Opfer wurde nur kurz gesetzt und nicht weiter gestärkt, was dazu führen könnte, dass er sich weniger nachhaltig unterstützt fühlt, als es möglich wäre. Bei Leon blieb es bei einer eher oberflächlichen Nachfrage, teils mit inhaltlichem Abweichen, was das Risiko birgt, dass er sich nicht zu einer ernsthaften Auseinandersetzung eingeladen fühlt. Bei Mira wurde ihre Idee zwar angesprochen, aber nicht ausdrücklich weitergetragen oder gemeinsam mit der Klasse aufgenommen, sodass sie den Eindruck gewinnen könnte, dass ihr Vorschlag eher eine Randbemerkung bleibt. Für einen nächsten Durchgang könnte daher hilfreich sein, den Blick noch klarer auf das Opfer zu richten, den potenziellen Täter bzw. Mitbeteiligten gezielter zu suchen und zu vertiefen und Miras Rolle als engagierte Mitschülerin bewusster zu stärken.

Callout 7: Transcript of the feedback for the criterion 'Target Group Focus'.

3.11. Text-to-Speech

As previously noted, the feedback is delivered via an AI-generated voice to provide an immersive and engaging experience. The technical implementation of this feature is contained within the FeedbackTTS.cs file.

For the TTS generation, the system uses OpenAI's Audio API. The model selected for the application is gpt-4o-mini-tts, which currently represents the most efficient and reliable option for real-time synthesis.

A standard request to the API requires the following four parameters:

- *Model:* Specified as gpt-4o-mini-tts.
- *Voice:* OpenAI provides 11 different vocal profiles, including *alloy*, *ash*, *ballad*, *coral*, *echo*, *fable*, *onyx*, *nova*, *sage*, *shimmer*, *and verse*. For the VR-Academy feedback system, the "nova" voice was selected.
- *Input:* The specific feedback text to be converted into speech. Consistent with the prompt construction, this string is sanitized using the method illustrated in Code Snippet 12.
- *Instructions:* This parameter allows for the customization of personality, tone, pronunciation, etc.

To ensure a supportive learning environment, a predefined instruction set titled "Friendly" was sourced from OpenAI.fm [20]. The specific instruction string is documented in Callout 8.

Voice Instructions (Old)

Affect/personality: A cheerful guide.

Tone: Friendly, clear, and reassuring, creating a calm atmosphere and making the listener feel confident and comfortable.

Pronunciation: Clear, articulate, and steady, ensuring each instruction is easily understood while maintaining a natural, conversational flow.

Pause: Brief, purposeful pauses after key instructions to allow time for the listener to process the information and follow along.

Emotion: Warm and supportive, conveying empathy and care, ensuring the listener feels guided and safe throughout the journey.

Callout 8: Previous voice instructions for the feedback texts.

Initial testing revealed that the default reading pace was quite slow compared to the length of the feedback segments. To improve the user experience, the voice instructions were modified to increase the speech rate while preserving the supportive and "friendly" characteristic. The refined and final instruction set is documented in Callout 9.

Voice Instructions (New)

Affect/personality: An energetic and cheerful guide.

Tone: Friendly, efficient, and reassuring, creating a positive atmosphere.

Pronunciation: Clear and fluid, ensuring each instruction is understood while maintaining a brisk, natural conversational flow.

Pace: Forward-moving and engaging, minimizing downtime between sentences to keep the listener attentive.

Emotion: Warm and supportive.

Callout 9: Final voice instructions for the feedback texts.

In contrast to the primary API call used for text generation, the TTS generation process utilizes a simplified error-handling mechanism. Because the audio component is considered a supplementary enhancement rather than a critical system requirement, the logic is limited to a single retry after a five-second delay. This ensures that even in the event of a persistent TTS failure, the user remains capable of advancing through the evaluation and reading the text-based feedback without obstruction.

3.12. Operational Overview of the Popup Manager

A comprehensive visualization of the logic within FeedbackPopupTextManager.cs is provided in Diagram 2.

In the diagram, it is assumed the process begins after the initial feedback text has been displayed. For conciseness, these preliminary loading steps (shown in Code Snippet 14) are omitted from the diagram. The user receives a total of five feedback texts: One for each of the four evaluation criteria and one short summary at the end.

The main process loop is initiated by user input. Upon a press of the “Next” button, any active audio playback is terminated, and the button is temporarily hidden to prevent premature state transitions before the next feedback audio item is ready.

Subsequently, a check determines if the current feedback is the last in the sequence. If true, the system proceeds to close the feedback. Otherwise, a secondary check identifies if the feedback is the penultimate item. In this case, the “Next” button is immediately re-displayed. This allows the user to advance to the final item without delay, as no further audio pre-fetching is necessary.

The system then exchanges the feedback texts and plays the pre-fetched audio corresponding to the current text. From this point on, the user views and listens to the feedback.

If this is not the final feedback item, an asynchronous TTS call is made to pre-fetch the audio for the next item. Should this operation fail, the system will attempt a single retry. Crucially, the “Next” button is always re-displayed after this stage, regardless of the TTS

call's success. This design ensures that user progression is not blocked by TTS service failures.

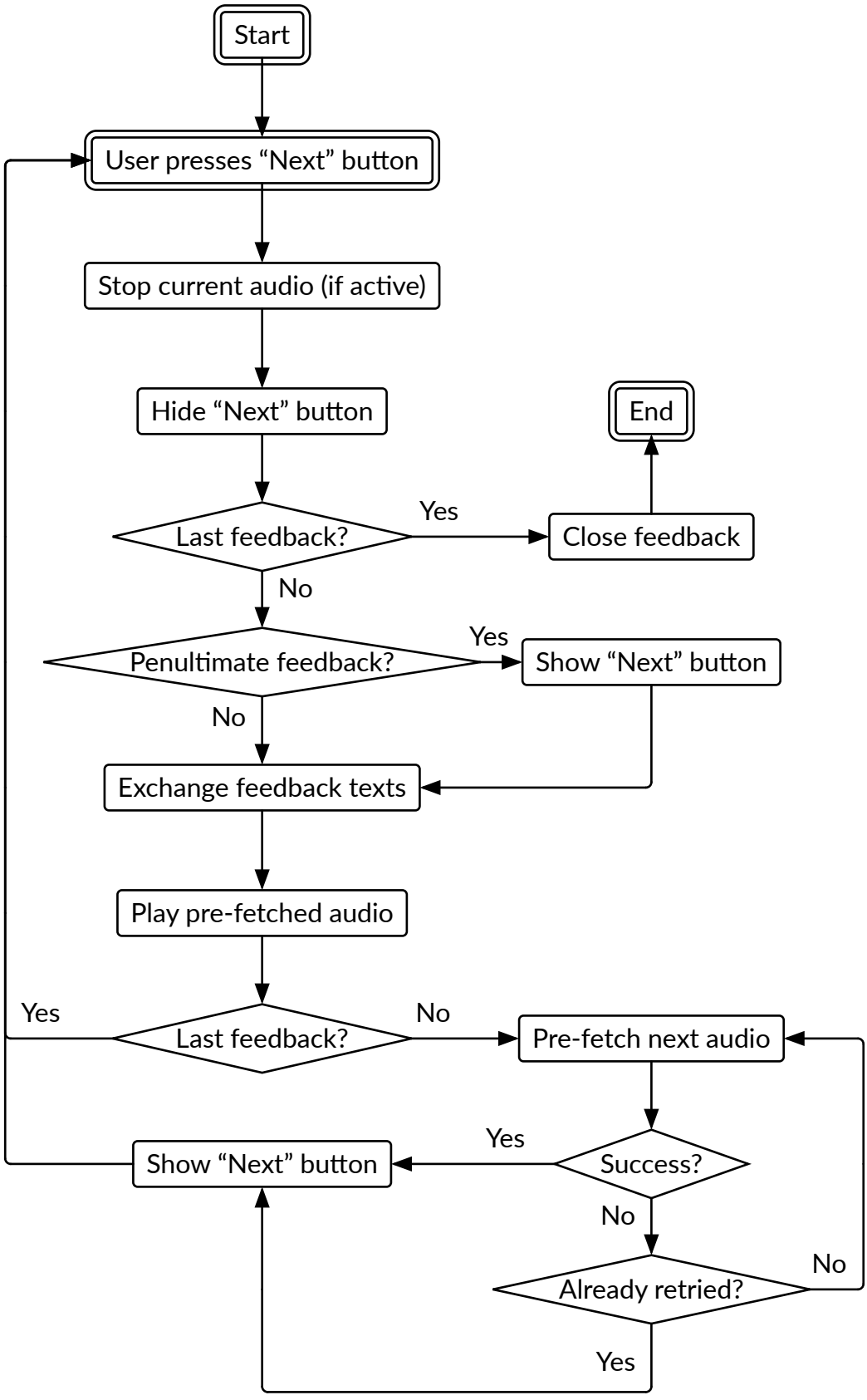

Diagram 2: Orchestrating input, feedback display, and TTS via FeedbackPopupTextManager.cs.

3.13. Prompt Engineering

The development of the optimal prompt for this application was an iterative process, characterized by continuous testing and refinement. The following sections delineate the evolutionary stages of this prompt, exploring how various models and configurations were evaluated to achieve the desired behavioral output and structural consistency from the LLM.

3.13.1. The Lack of Didactic Feedback

The initial prompt, documented in Appendix B, was deployed within the application to generate an initial set of feedback samples. The test scenario involved brief interactions with each individual student and a short address to the entire class via the broadcast system. For this phase, the GPT-5 Nano model was used. This variant represents the most latency-optimized and cost-efficient version of the GPT-5 family. Following the generation of these samples, the outputs were evaluated by project stakeholders during regular review sessions.

The initial prompt was designed to capture a fundamental overview of user performance. While it succeeded in providing descriptive accounts of the session, it lacked a rigorous didactic analysis of the user's pedagogical choices. Furthermore, the model failed to address or provide guidance on how the user could manage instances of discrimination within the classroom.

To address these limitations, the following refinements were implemented:

- The instructions were pivoted to prioritize pedagogical value and actionable didactic feedback.
- Each of the four evaluative criteria was required to address three specific dimensions:
  1. Identifying effective pedagogical actions.
  2. Pinpointing specific areas of weakness.
  3. Providing a concrete practical recommendation (with its underlying educational reason).
- The final summary was instructed to culminate in a single, "deciding" piece of advice for the teacher to implement in future sessions.

To ensure the "practical recommendation" met the desired quality standards, excerpts from the initial feedback concept were integrated into the prompt. This served as a stylistic template, dictating the professional and constructive tone the LLM should adopt.

3.13.2. The Necessity for Descriptive Feedback

The second iteration of the prompt is documented in Appendix C, with the specific modifications highlighted in green. As in the previous phase, feedback was generated following a short playthrough. However, this evaluation expanded to include both GPT-5 Nano and the standard GPT-5, which, as of October 2025, served as OpenAI's flagship model. This dual-model approach was designed to facilitate a comparative analysis,

determining whether the increased operational costs of the flagship model were justified by a significant improvement in feedback quality over the more economical Nano variant.

The results of this iteration revealed that the generated feedback tended to be overly judgmental, explicitly labeling player actions as good or bad. It was determined that the output should instead be more objective and descriptive. Furthermore, the analysis focused on the perceived emotional states of the students rather than the specific pedagogical interventions of the player. Finally, the feedback for individual students remained superficial, lacking the depth for meaningful reflection.

To mitigate these issues, the decision was made to leverage one-shot prompting. As established in Section 2.8.2, providing the LLM with a concrete demonstration can significantly improve output consistency and style. Consequently, a comprehensive feedback text from the original theoretical concept (referenced in Section 3.3) was integrated directly into the prompt to serve as a high-quality stylistic and structural template.

In summary, the third iteration involved a restructuring of the prompt instructions to ensure a higher standard of objectivity and depth. The following modifications were implemented:

- The feedback tone was pivoted from a judgmental “good/bad” dichotomy to a more objective, descriptive analysis.
- Each criterion was reorganized to address three specific educational perspectives:
  1. Action: A summary of one or multiple interventions made by the teacher.
  2. Positive Implication: An analysis of the possible pedagogical benefits or positive outcomes resulting from these actions.
  3. Downside/Risk: An exploration of potential risks, negative side effects, or pedagogical conflicts of interest associated with the chosen approach.
- To enforce this three-point structure, a concrete demonstration was integrated into the prompt instructions, as shown in Callout 10. This example illustrates the expected output format using a hypothetical interaction with the student David.
- Instructions were added to ensure that the LLM uses cautious, non-definitive language (e.g., “He may feel” or “He could feel”) when interpreting student emotional states.
- The LLM was directed to finalize the analysis for each individual student before transitioning to the next.
- Word limits were removed to allow the model sufficient space to provide in-depth, nuanced feedback.

Drawing on the prompt engineering strategies proposed by Indran et al. [14], as discussed in Section 2.8.1, the prompt’s instructional architecture was transitioned from a paragraph-based format to a sequentially numbered list.

Furthermore, the implementation of length constraints was a subject of considerable deliberation. Ultimately, it was decided to remove any explicit word or character limits. This choice ensures that the generated feedback remains sufficiently comprehensive to offer genuine pedagogical value. Moreover, removing these constraints provides a valuable opportunity to evaluate the behavior and verbosity of LLMs when tasked with generating complex educational assessments.

Demonstration (in German)

Wenn in Konfliktgesprächen der Fokus auf die betroffene(n) Person(en) gelegt wird, kann dies ein wichtiges Signal an die gesamte Klasse und insbesondere an die Betroffenen senden: Ihre Perspektive wird wahrgenommen, ihre Erfahrungen werden ernst genommen. In angespannten Situationen kann dies ein Gefühl von Anerkennung und Wertschätzung vermitteln. Gleichzeitig besteht jedoch die Gefahr, dass die eigentliche diskriminierende Handlung und das Verhalten der Täter*innen in den Hintergrund treten und nicht ausreichend reflektiert oder aufgearbeitet werden.

Callout 10: One-shot-prompting: Usage of a demonstration in the prompt.

3.13.3. An Improved Prompt Version

The third iteration of the prompt is provided in Appendix D, with the newest modifications highlighted in green. Feedback was again generated using both GPT-5 and GPT-5 Nano and reviewed by project stakeholders.

The resonance from this iteration was overwhelmingly positive. The resulting structure closely mirrored the original pedagogical concept, successfully illuminating teacher actions alongside their positive and negative implications.

However, certain minor refinements were identified. The requirement for cautious phrasing regarding student emotions was not consistently applied across all segments. Additionally, the analysis occasionally transitioned between students before a comprehensive profile was completed. There was also a concern regarding accessibility; terminology such as “bystander behavior”, “explicit validation”, and “modeled empathy” was deemed potentially too specialized for some users. Consequently, the prompt was adjusted to utilize more common pedagogical terminology.

Furthermore, the model exhibited a tendency to “hallucinate” suggestions for improvement, that is, proposing pedagogical actions that were not grounded in the specific unachieved flags. While these suggestions might be pedagogically sound, their accuracy regarding the actual simulation could not be guaranteed. To ensure reliability, the feedback was restricted to the scope of the pre-defined flag system.

Finally, it was concluded that the output quality of GPT-5 Nano was insufficient for the system’s requirements. While GPT-5 was the preferred choice, the team decided to evaluate GPT-5 Mini, which is a faster and more cost-effective variant, to determine if it could provide a balance between performance and quality.

The following refinements were implemented in the prompt:

- Reinforced the requirement to identify potential student emotions using non-definitive phrasing, such as “He may feel.” or “He could feel.”
- Mandated that each student’s evaluation be completed in full before the LLM transitions to the next student.
- Explicitly prohibited jargon like “bystander behavior”, requiring the use of more pedagogical terms instead.

- Restricted suggestions for improvement strictly to the scope of achievable flags.
- Formally defined the LLM's persona as a "Coach for didactics and classroom management".
- Transitioned to a "one paragraph per point" output format to enhance the readability of the feedback for the user.

3.13.4. The Final Prompt Version

The fourth and final iteration of the prompt is documented in Appendix E, with the latest refinements highlighted in green. For this concluding evaluation, feedback was generated using both GPT-5 and GPT-5 Mini.

The feedback received for these texts was overwhelmingly positive, and no further areas for improvement were identified. While the outputs from GPT-5 Mini were slightly less sophisticated than those produced by the flagship GPT-5 model, they were deemed sufficiently high-quality for integration into the application. The selection of GPT-5 Mini was driven by its economic and operational advantages, as it is both more cost-effective and faster than GPT-5. This ensures greater scalability for future use. Should GPT-5 Mini demonstrate any performance weaknesses during the testing phase, the system remains capable of reverting to GPT-5, or even switching to a newer model.

4. Study Design

This chapter describes the design of the user study conducted to evaluate the performance of the implemented feedback system. Furthermore, it outlines the methodology used to compare the AI-generated evaluations against traditional feedback provided by human experts.

4.1. General

The quality of the implemented feedback system was evaluated through a structured user study. Two different participant groups were recruited: pre-service teachers and non-teachers. Participants engaged with the first VR-Academy scenario, which focuses on addressing anti-Semitism. Following the simulation, each participant was assessed by both the AI-assisted system and a human evaluator. Both feedback types were structured identically: each addressed the four specific evaluation criteria and provided a brief summary at the conclusion.

The study employed a within-subjects design, meaning every participant received feedback from both feedback sources. This approach allowed for a comparative analysis of the feedbacks. To mitigate potential order effects, participants were divided into two groups: one received the AI feedback first, while the other began with the human evaluation.

Data collection followed a mixed-methods approach using Google Forms, incorporating both quantitative and qualitative measures:

- Quantitative: Participants rated specific feedback dimensions on a five-point Likert scale and answered closed-ended, multiple-choice questions.
- Qualitative: Open-ended questions allowed participants to provide insights into their experience.

As detailed in Section 3.8, the GPT-5.1 model was utilized for the study. As OpenAI's flagship model as of November 2025, it is an ideal candidate for competing against professional human pedagogical feedback. To ensure a high benchmark for pedagogical expertise, the human feedback was provided by a post-doctoral researcher at the professorship for Islamic Religious Education and Subject Didactics.

4.2. Data Collection

The following provides a summarization of the questions used across the Google Form surveys. To maintain anonymity and organize data, each participant was assigned a unique ID. The prefix indicated the assessment order and the participant's background (pre-

service teacher or non-teacher). For example, a participant with the ID “A3” received AI feedback first, whereas “B3” began with human evaluation. Additionally, every survey included an “Other comments” field to capture any insights not addressed by specific questions. Responses on the five-point Likert scale ranged from (1) “I strongly disagree” to (5) “I completely agree”.

While most questions consist of ad-hoc items developed specifically for this study, one exception is discussed in a subsequent section.

The initial survey questions are detailed in Callout 11. This questionnaire was administered prior to the VR session, beginning with a declaration of consent and a privacy policy. The next sections collected demographic data and assessed participants’ prior experience with VR technology, educational simulations, and video games. Administering this survey beforehand was crucial to establishing an unbiased baseline.

Survey: Demographics and Previous Experience

Age

Gender

Study program

Semester

Experience with Virtual Reality (VR): Scale in hours: 0, <1, <10, <100, 100+

Experience with simulations in an educational context: Scale in hours: 0, <1, <10, <100, 100+

Experience with video games: Scale in hours: 0, <1, <10, <100, 100+

Other comments

Callout 11: Pre-study survey focusing on demographics and prior experience.

The survey questions for blocks B and C are detailed in Callout 12. This specific questionnaire was administered both after the AI-assisted feedback and the human evaluation, regardless of the sequence. Using identical metrics for both interventions ensured methodological consistency, facilitating a valid comparison between the two feedbacks.

The first segment of the survey evaluated the feedback across six core dimensions: comprehensibility, concreteness, accuracy, tone, structure, and length.

Following these descriptors, four items were presented to assess reflection, improvement, effectiveness, and overall utility. These items were adapted from the TAM by Davis [15] to measure the Perceived Usefulness of the feedback, as discussed in Section 2.9.

All quantitative items were rated on a five-point Likert scale. To conclude each block, two qualitative open-ended questions were included, prompting participants to identify the most and least useful aspects of the respective feedback.

Survey: Evaluation of the AI Feedback System and Human Feedback

Comprehensibility: The feedback was clear and easy to understand. (Scale 1-5)

Concreteness: The feedback gave me concrete suggestions on what I could do differently. (Scale 1-5)

Accuracy: The feedback has accurately analyzed my actions and their possible consequences in the scenario. (Scale 1-5)

Tone: I perceived the feedback as supportive coaching, not as a judgmental evaluation. (Scale 1-5)

Structure (Balance): The balanced nature of the feedback (positive and negative implications) was useful for seeing different perspectives. (Scale 1-5)

Length: How do you rate the overall length of the feedback? (Scale: Way too short (1) to Way too long (5))

Reflection: The feedback was helpful for reflecting on my own behavior. (Scale 1-5)

Improvement: The feedback will help me improve my skills for similar situations in the future. (Scale 1-5)

Effectiveness: The feedback increases my effectiveness in learning how to handle such situations. (Scale 1-5)

Overall Utility: Overall, I find this feedback very useful for my training. (Scale 1-5)

What were the most useful aspects or specific phrases of the feedback? (Open question)

What was least useful or was missing? (e.g., unclear, inaccurate, too general, something important was ignored) (Open question)

Other comments

Callout 12: Survey evaluating the AI feedback and human feedback.

While the survey questions remained identical for both feedback methods to ensure comparability, the AI evaluation block included an additional sub-section regarding the AI-generated voice. These additional items are illustrated in Callout 13.

The sub-section commenced with an inquiry regarding the overall pleasantness of the voice. Subsequent items evaluated clarity, speaking speed, and sound. This block concluded with an open-ended question, allowing participants to provide further remarks.

Survey: Evaluation of the AI-Generated Voice

How do you rate the AI-generated voice of feedback overall? (Scale: Very unpleasant (1) to Very pleasant (5))

Clarity: The voice was clear and distinct. (Scale 1-5)

Speaking speed: The pace was appropriate. (Scale 1-5)

Sound: The voice sounded human/natural. (Scale 1-5)

Did you have any comments about the voice (e.g., emphasis, tone)? (Open question)

Callout 13: Survey evaluating the AI-generated voice.

The user study concluded with a final comparative survey, as illustrated in Callout 14.

Survey: Comparison and Acceptance of AI vs. humans

I felt comfortable being evaluated by an AI. (Scale 1-5)

In direct comparison, which feedback did you perceive as... (Scale: Clearly AI (1) to Clearly Human (5))

...more helpful?

...more objective?

...more motivating?

...more detailed?

What kind of feedback would you prefer for your future training? Options: AI feedback only, Human feedback only, A combination of both, Neither.

Why? (Open question)

Other comments

Callout 14: Final survey comparing AI vs. human feedback.

The final survey focused on the direct comparison and acceptance of AI versus human evaluators. Participants were asked to rate their comfort level regarding being evaluated by an AI. Furthermore, the survey included comparative items to determine which feedback source was perceived as more helpful, objective, motivating, and detailed. The evaluation ended in a final question regarding which method participants would prefer for their future training, accompanied by a qualitative justification for their choice.

4.3. Study Procedure

A comprehensive visualization of the experimental procedure is provided in the flowchart in Diagram 3.

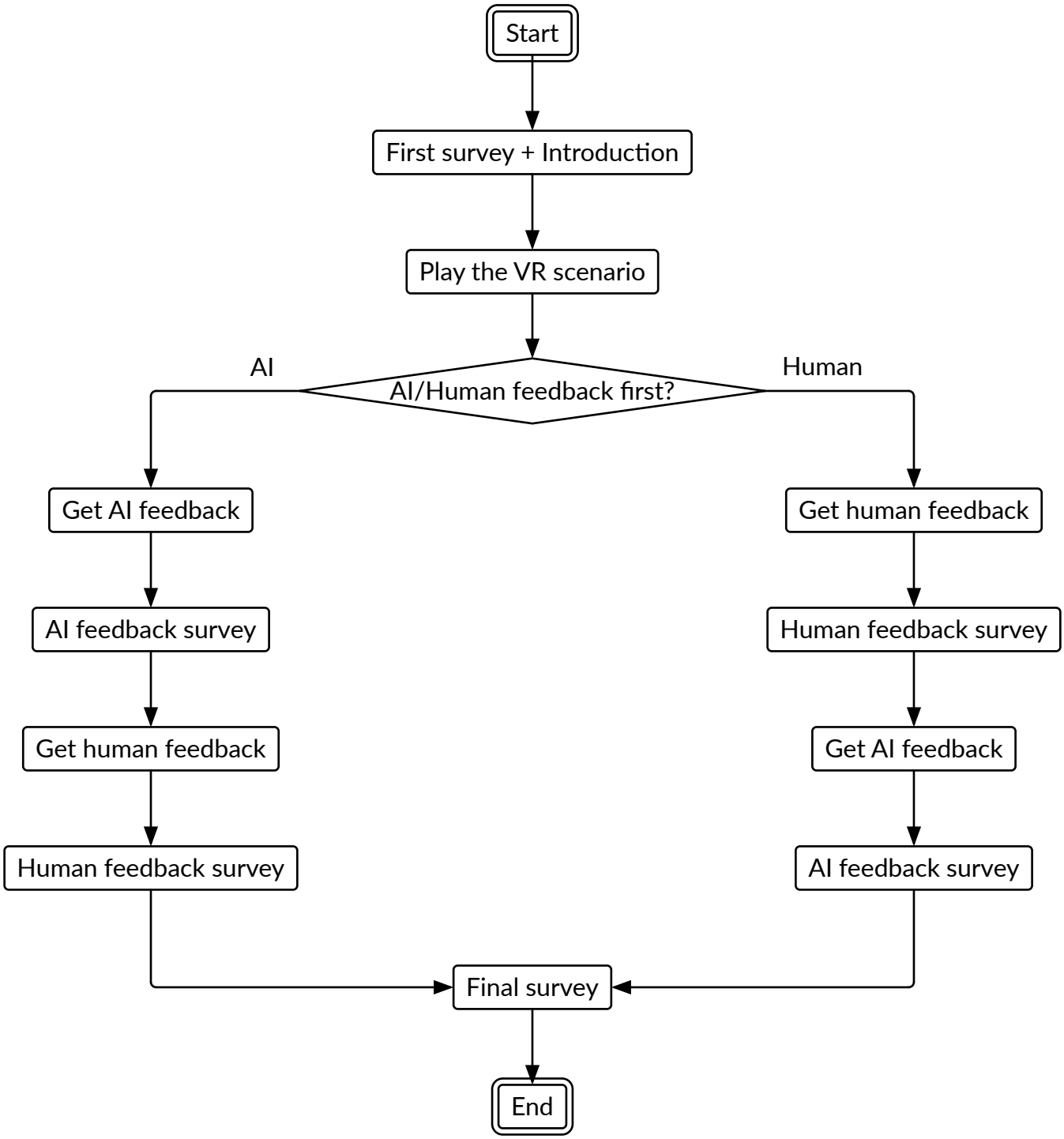

Diagram 3: Overview of the study procedure.

The study commences with each participant completing the initial survey. To ensure a standardized level of preparation, an introduction was provided featuring three specific application screenshots: the first depicted the classroom environment to identify interactable students, the second illustrated the chat interface to explain the communication mechanics, and the third demonstrated the procedure for concluding the scenario. Following this briefing, participants engaged with the VR simulation.

The procedural divergence is illustrated in the subsequent paths: the left path delineates the sequence for participants receiving AI-assisted feedback first, followed by a survey

and the human evaluation. Conversely, the right path outlines the inverse order for the second group. Both pathways converge at the end of the study, concluding with the final survey.

5. Results

This chapter presents the findings from the user study and the corresponding statistical analyses.

5.1. Participants

A total of 14 participants were recruited for this study ($N = 14$). The sample was divided into two equal groups: pre-service teachers ($n = 7$) and a control group of non-teachers ($n = 7$). Unless otherwise stated, these group sizes remain constant throughout the analysis. Table 1 shows a summary of demographic data for the participants.

Variable	Pre-Service Teachers	Non-Teachers	Total
Age (Mean)	28.4	24.7	26.6
Male	6	7	13
Female	1	0	1

Table 1: Participant demographics.

Within the group of pre-service teachers, the age of participants ranged from 21 to 36 years. The non-teacher group exhibited a narrower age distribution, with participants aged between 21 and 28.

The pre-service teachers represented a diverse academic spectrum, majoring in subjects including Islamic Religious Education, German, English, Mathematics, History, Biology, Physical Education, and Pedagogy. In contrast, the non-teacher participants possessed backgrounds in fields such as Software Engineering, Mechanical Engineering, and Warehouse Logistics.

Although data regarding the current semester was collected, the survey design did not differentiate between semesters enrolled in the current program (Fachsemester) and total semesters at the university (Hochschulsemester). Due to this varying interpretation by participants, the resulting dataset lacked the consistency required for valid statistics and was therefore excluded.

The participants' prior experience with VR is illustrated in Figure 10. The majority of the pre-service teachers reported having no prior experience. Conversely, all participants in the non-teacher group had utilized VR at least once (<1 hour), with their cumulative hours of experience being broadly distributed across the scale.

The level of participant experience with simulations in educational contexts is detailed in Figure 11. The majority of participants across both groups possessed little to no prior

exposure to such simulations. A single pre-service teacher represented an outlier in this category, reporting a more significant background of less than 100 hours of experience.

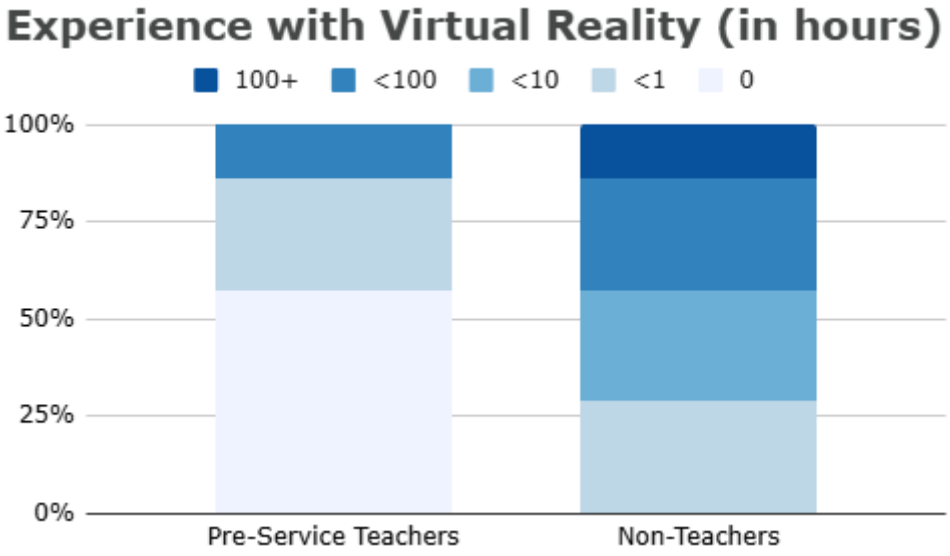

Figure 10: Participants' experience with Virtual Reality.

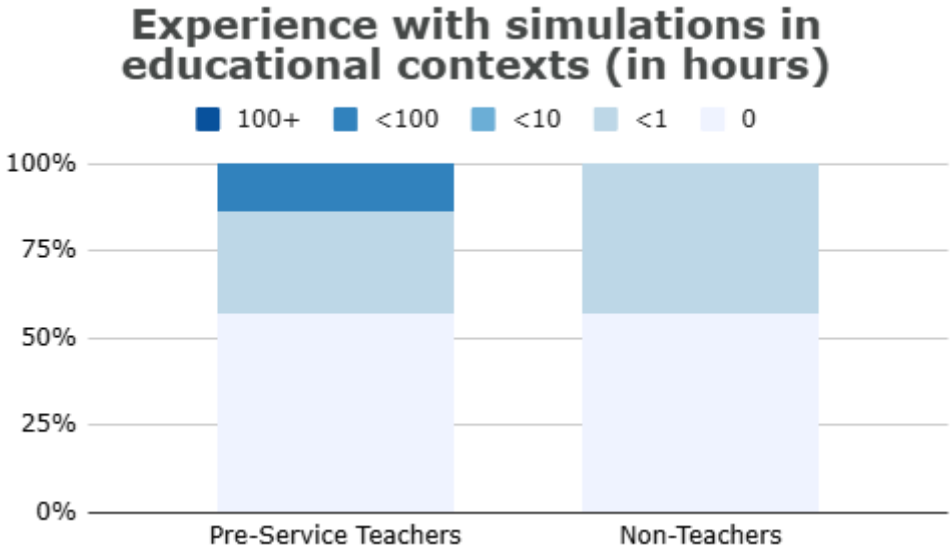

Figure 11: Participants' experience with simulations in educational contexts.

In contrast, Figure 12 illustrates a higher degree of familiarity with video games. Both groups had individuals with extensive gaming experience, exceeding 100 hours. Within the pre-service teacher group, one participant reported no prior experience, whereas all non-teachers had engaged with video games for at least some period (<10 hours).

Regarding qualitative remarks, one pre-service teacher specified that they own a PlayStation VR2 which is a specialized VR peripheral developed exclusively for the PlayStation 5 console.

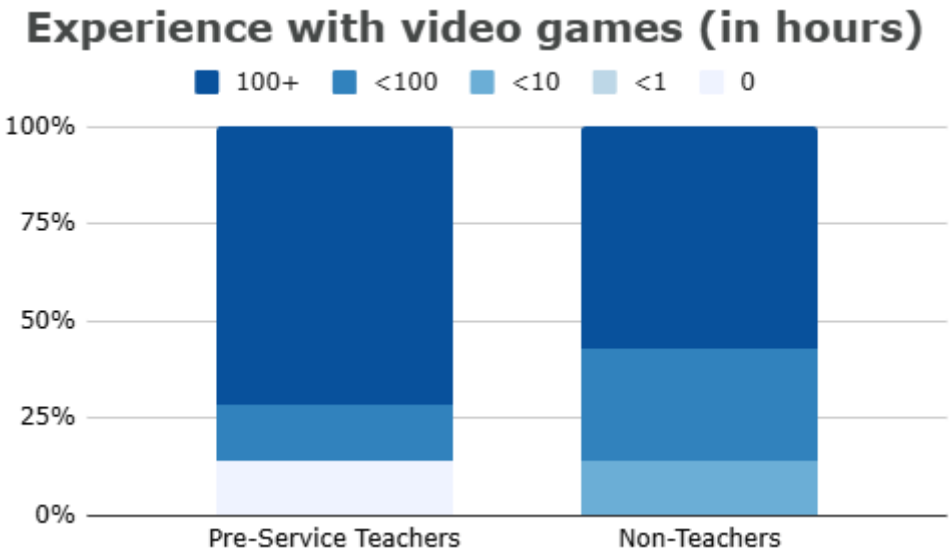

Figure 12: Participants' experience with video games.

5.2. AI and Human Ratings

The following sections present the results of the surveys administered to participants immediately following each feedback.

5.2.1. Quantitative Results

Table 2 illustrates the ratings for each feedback, disaggregated by participant group (pre-service teachers vs. non-teachers) and feedback source (AI vs. human). Dimensions are abbreviated for conciseness; comprehensive definitions are provided in Callout 12. Results are reported as Mean [Median]. While most dimensions utilized a 5-point Likert scale where 5 represents the highest rating, the "Length" dimension follows a central tendency where 3 is considered the optimal score.

In aggregate, human feedback yielded higher mean ratings across most dimensions compared to AI feedback. The AI's lowest mean score was 3.71 for "Accuracy" (marked in red), as reported by pre-service teachers, while its highest rating reached 4.86 in "Comprehensibility" (marked in green), also within the pre-service teacher cohort.

Conversely, human feedback received its lowest mean ratings in "Reflection" (4.29) and "Length" (3.71), both from pre-service teachers (marked in red). Its highest mean score was a 5.00 in "Concreteness" (marked in green), awarded by the non-teacher group. The AI feedback outperformed human feedback in mean ratings only within the pre-service teacher group for the dimensions of "Comprehensibility" and "Reflection."

Regarding the distribution of scores, human feedback received no ratings of 2 or lower across any dimension. Similarly, AI feedback received no ratings of 1.

Dimension	Pre-Service Teachers		Non-Teachers	
	AI	Human	AI	Human
Comprehensibility	4.86 [5]	4.43 [5]	4.29 [5]	4.86 [5]
Concreteness	4.43 [5]	4.57 [5]	4.14 [4]	5.00 [5]
Accuracy	3.71 [5]	4.57 [5]	4.29 [5]	4.71 [5]
Tone	4.57 [5]	4.71 [5]	3.86 [4]	4.86 [5]
Structure	4.71 [5]	4.86 [5]	3.86 [4]	4.86 [5]
Length (Optimal: 3)	3.86 [4]	3.71 [3]	3.86 [4]	3.57 [3]
Reflection	4.43 [4]	4.29 [4]	4.00 [4]	4.86 [5]
Improvement	4.14 [4]	4.86 [5]	3.86 [4]	4.71 [5]
Effectiveness	3.86 [4]	4.43 [5]	3.86 [4]	4.57 [5]
Overall Utility	4.14 [4]	4.86 [5]	4.29 [5]	4.71 [5]

Table 2: Comparison of AI and human feedback ratings across both groups (Mean [Median])

Notably, one participant in the non-teacher group evaluated the AI feedback more critically than the average. This outlier substantially influenced the mean ratings for AI feedback. However, the median values demonstrate that AI feedback consistently maintained a score of at least 4. A similar trend is observed in the human feedback medians, with the exception of "Length", where the median aligned with the optimal value of 3.

5.2.2. Perceived Usefulness

As previously noted, the final four dimensions "Reflection", "Improvement", "Effectiveness", and "Overall Utility" were adapted from Davis [15]. Consequently, the scores for these items were aggregated to determine an overall rating for the Perceived Usefulness of both feedback types.

The summarized scores, reported as Mean [Median], are as follows:

- Pre-Service Teachers: AI: 4.14 [4.00]; Human: 4.61 [4.75].
- Non-Teachers: AI: 4.00 [4.25]; Human: 4.71 [4.75].

In both participant groups, the Perceived Usefulness ratings were higher for human feedback than for AI feedback. These aggregated values, rather than individual dimension scores, will be utilized for subsequent statistical testing.

5.2.3. Qualitative Results

The following section presents a qualitative summary of the responses to the open-ended questions, aggregating feedback from both participant groups.

Positive attributes of the AI feedback:

Participants consistently noted that the AI feedback facilitated a deeper understanding of complex scenarios. Multiple respondents highlighted that the feedback offered a balanced view by outlining both the positive implications and downsides of their actions. This approach reportedly exposed blind spots and offered perspectives the players had not previously considered. Furthermore, the feedback successfully fostered empathy by illustrating how specific actions might influence student feelings.

The feedback was praised for its concreteness. Participants valued the explicit use of certain words (e.g. Antisemitism) and the inclusion of specific instructions, such as checking on a victim's well-being or defining the scope of a class project, rather than vague advice. Respondents also noted that the system was helpful in identifying gaps in their performance, specifically pointing out topics they had neglected to address.

Participants reported that the AI correctly identified their actions for the most part. Additionally, the simultaneous presentation of text and voice was cited as a key usability feature, allowing users to process the information via audio while also having the text available.

Negative attributes of the AI feedback:

A primary concern among participants was the system's occasional failure to recognize the context or constraints of the simulation. Participants noted that the AI criticized them for failing to achieve specific goals, such as addressing the class or concluding a conversation, even when a virtual student had refused to speak further.

Additionally, issues with the classification of player statements and TTS recognition sometimes led to feedback that users felt was incorrect or strange. For instance, one participant noted that the system misinterpreted their annoyance with the student Mira as an attempt to block her statements leading to irrelevant advice.

Participants frequently cited the length of the feedback as a negative factor. The summaries for each criteria were described as too long, often containing redundant information when outlining both positive implications and downsides. Users suggested that the system should prioritize focus more heavily on concrete suggestions for improvement rather than providing exhaustive retrospective analysis.

In several instances, participants challenged specific pieces of advice by the AI. For example, suggestions to discuss affected individuals in front of the whole class were deemed questionable due to the potential for making victims feel uncomfortable. Similarly, players disagreed with the feedback recommending the immediate removal of writing from the blackboard, noting that their intent was to preserve it as evidence for the school principal.

Other remarks about the AI feedback:

One participant noted that the feedback criticized the lack of mentioning "Antisemitism", even though the player specifically mentioned "terrible acts in our history", which exactly implies this term.

Positive attributes of the human feedback:

Participants highlighted the constructive nature of the human feedback. Specifically, they noted that it prioritized concrete suggestions for improvement and future behavior rather than focusing primarily on wrong actions. Respondents reported that the feedback clearly differentiated between successful actions and areas for growth, offering specific guidance on how to better approach the topic of discrimination in the future.

A recurring point in the responses was the focus on interpersonal dynamics. Participants stated that the feedback illustrated potential student feelings and offered guidance on how to show empathy toward both victims and perpetrators. Additionally, respondents noted that the feedback helped them adopt a professional teacher's role, for instance, by suggesting strategies for addressing the topic with the whole class rather than solely relying on separate confrontations.

Participants also valued the range of strategies presented. They reported that the feedback outlined both the pedagogical advantages and disadvantages of their choices. Notably, respondents appreciated that alternative approaches were provided even for actions deemed good, not just for bad ones. Specific scenarios, such as the timing of removing writings from the blackboard, were cited by participants as helpful examples that showed multiple potential solutions to facilitate reflection.

Negative attributes of the human feedback:

Some participants noted limitations regarding the depth of the human feedback. One respondent remarked that the feedback did not provide concrete options for action for every single criterion. Additionally, criticism was raised regarding the scope of the evaluation. Specifically, participants noted a lack of feedback concerning their line of thoughts during the decision-making, a limitation they observed in the AI feedback as well. One participant also expressed dissatisfaction with the discussion regarding the different approaches to the problem.

When comparing the two feedback modalities, some participants perceived the human feedback as offering less specific criticism than the AI regarding the consequences for individual students. Respondents noted that the human feedback tended to focus on the consequences for the general classroom climate rather than the direct impact on specific students. Consequently, one participant explicitly stated that the AI feedback was superior in demonstrating the specific consequences that would arise if the player followed their own actions.

Other remarks about the AI feedback:

One participant suggested that the body language should also be taken into account for the feedback. For example, the player could be standing in front of the class, but lowering himself to eye level when talking to David.

5.3. AI Voice

This section presents the results concerning the participant perceptions of the AI-generated voice used for the feedback.

5.3.1. Quantitative Results

Table 3 presents the ratings from both participant groups regarding the AI-generated voice. As with the previous table, dimensions are abbreviated for conciseness; detailed definitions are available in Callout 13. Data are reported as Mean [Median], using a 5-point Likert scale where 5 represents the optimal score and 1 the lowest.

Dimension	Pre-Service Teachers	Non-Teachers
Overall pleasure	3.57 [4]	4.00 [5]
Clarity	4.57 [5]	4.43 [5]
Speaking speed	3.57 [4]	4.29 [5]
Sound	3.14 [3]	3.43 [3]

Table 3: Participant ratings of the AI-generated feedback voice.

The non-teacher group provided higher ratings than the pre-service teachers in three of the four dimensions. The highest rating was recorded for “Clarity” (4.57) by the pre-service teacher group (marked in green). Conversely, the lowest rating was observed in the “Sound” dimension (3.14), also reported by the pre-service teachers (marked in red).

5.3.2. Qualitative Results

Participants were provided the opportunity to offer remarks regarding the AI voice, which are summarized below:

Several participants commented on the tempo of the voice. Respondents noted that the reading speed was too slow, which, combined with the overall length of the feedback, made it difficult to maintain concentration.

Opinions regarding the prosody and quality of the voice were mixed. On the positive side, some participants observed that the intonation was generally good and that the voice placed emphasis in the right spots. However, others described the voice as monotonous, particularly given the length of the text or noted that the pitch was sometimes too high. One participant described the overall tone as didactic. Additionally, remarks were made regarding consistency: one user noted that the voice changed at times, making it difficult to comprehend, while another suggested that the emphasis on certain words could have been better.

5.4. Final Survey

The following section presents the findings from the final survey administered at the end of the study.

5.4.1. Comfort

Figure 13 illustrates the participants' self-reported comfort levels regarding evaluation by an AI.

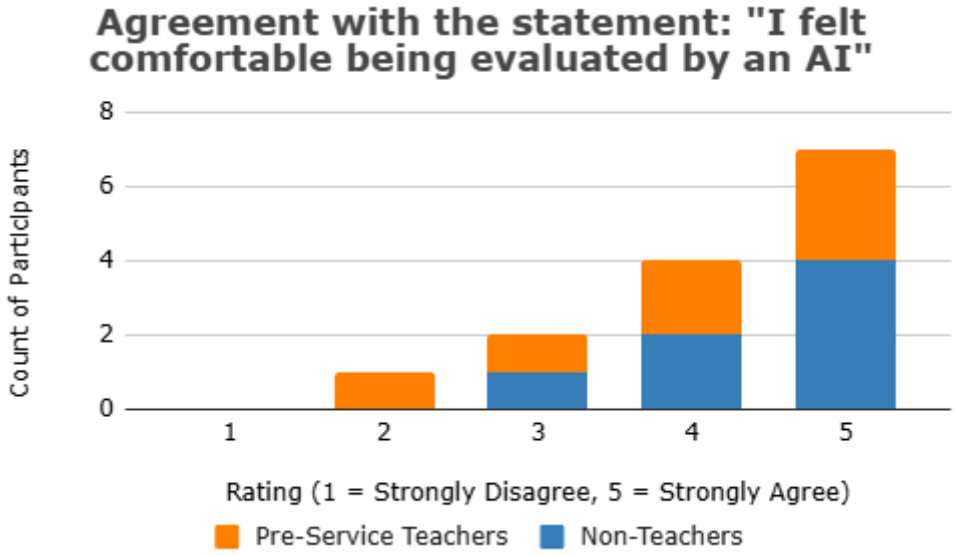

Figure 13: Participant comfort levels regarding AI evaluation.

The majority of respondents in both groups reported a comfort level of 3 or higher. Only one pre-service teacher provided a rating of 2, and no participants selected the "strongly disagree" (1) option.

The descriptive statistics for these responses, reported as Mean [Median], are as follows:

- Pre-Service Teachers: 4.00 [4]
- Non-Teachers: 4.43 [5]

5.4.2. Helpfulness, Objectivity, Motivation, and Detail

Figure 14 illustrates the preferences of pre-service teachers regarding feedback helpfulness, objectivity, motivation, and level of detail.

The majority of pre-service teachers identified "Clearly Human" or "Somewhat Human" feedback as the more helpful. Regarding objectivity, the distribution shifted slightly toward the AI end of the spectrum, with three participants selecting "Somewhat AI." In terms of motivation, six out of seven pre-service teachers favored human feedback ("Clearly Human" or "Somewhat Human"). Finally, most pre-service teachers categorized the AI feedback as more detailed ("Clearly AI" or "Somewhat AI"), while the remaining participants rated both feedback types as equal.

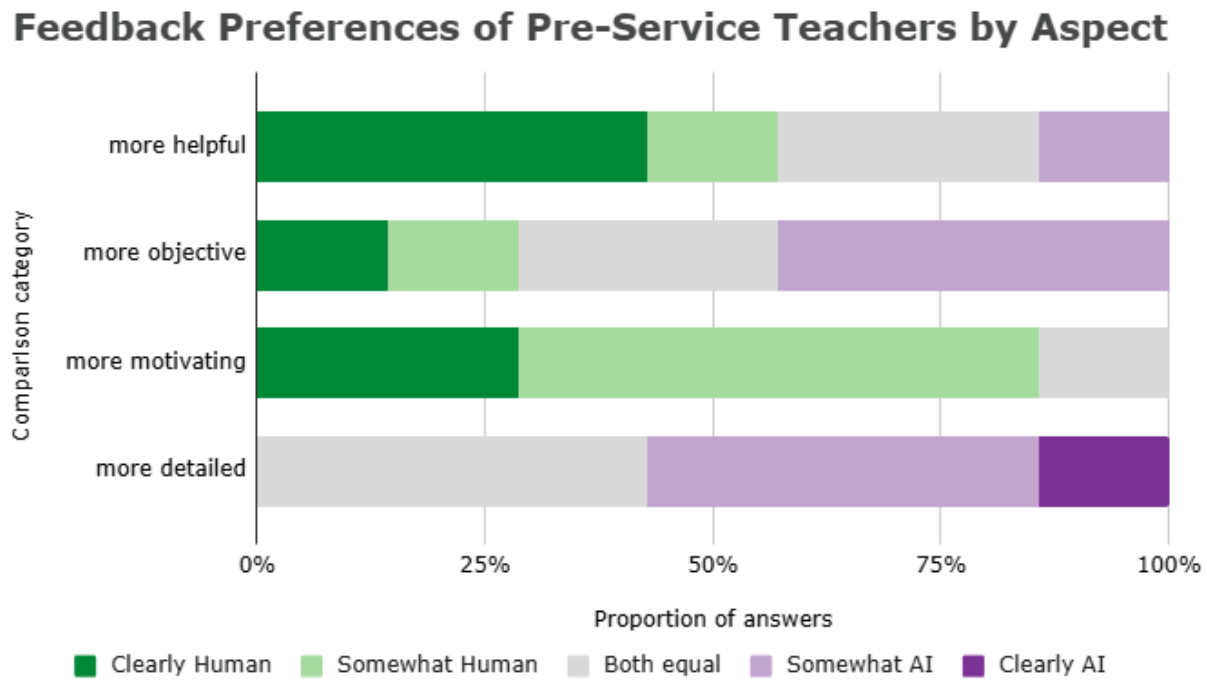

Figure 14: Pre-service teacher preferences for AI vs. Human feedback stratified by aspect.

Figure 15 presents the corresponding results for the non-teacher cohort.

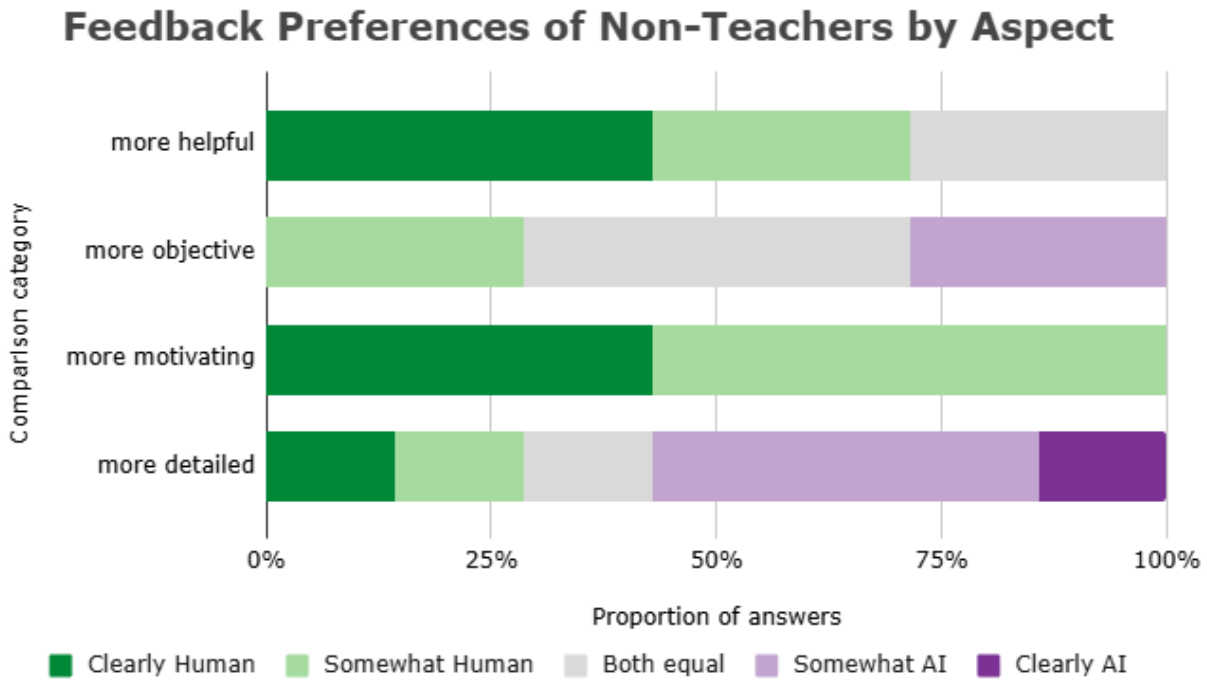

Figure 15: Non-teacher preferences for AI vs. Human feedback stratified by aspect.

Consistent with the pre-service teacher group, the majority of non-teachers identified "Clearly Human" or "Somewhat Human" as the more helpful feedback. Regarding objectivity, the distribution was evenly balanced, with two respondents each selecting "Somewhat Human" and "Somewhat AI", while the remaining participants reported

both as equal. Notably, all seven non-teachers selected “Clearly Human” or “Somewhat Human” as being more motivating. In terms of detail, the distribution trended toward the AI end of the spectrum, with the majority of non-teachers selecting “Clearly AI” or “Somewhat AI.”

Across the total sample of 14 participants, none selected “Clearly AI” or “Somewhat AI” regarding motivation. Overall, the “Clearly AI” designation was utilized only twice, specifically in reference to the level of detail provided in the feedback.

5.4.3. Future Preference

Participants were surveyed regarding their preferred feedback source for future training. All 14 participants (100%) indicated a preference for a hybrid approach, incorporating a combination of both AI and human feedback.

A summary of their justifications can be found below:

A recurring theme in the justifications was the view that the two feedback types complement each other’s strengths and weaknesses. Participants described the AI as offering a detailed and objective perspective, which they contrasted with the motivating and subjective perspective of the human. Respondents argued that this combination created an optimal balance for learning. One participant summarized this by noting that more is always better, suggesting that the AI covers aspects that humans might overlook, leading to both qualitative and quantitative advantages.

Participants drew a sharp distinction between the analytical capabilities of the AI and the contextual understanding of the human. The AI was praised for its ability to recall specific events and details and for identifying patterns. However, respondents noted that humans were better suited to comprehending intentions, recognizing system errors, and understanding complex emotions and the general classroom climate. One participant specifically noted that the AI sometimes interpreted emotions too hypersensitively compared to real-world practice.

Participants described specific use cases for each feedback. The AI was highlighted as a fast and easy tool for practicing alone. However, respondents noted that for this workflow to be effective, humans should remain available to discuss important details later. Regarding the nature of the interaction, opinions varied: while one participant described the AI as soulless, another appreciated that it did not generate a feeling of being judged. Conversely, one respondent reported that the AI rated decisions too positively, whereas the human feedback offered more diverse approaches.

Other remarks:

One participant mentioned that he got a racing heart from nervousness when he received the human feedback. During the AI feedback, he simply accepted it and used it for reflection.

5.5. Statistical Analysis

To identify potential significant differences, two statistical analyses were performed on the data sets.

5.5.1. Mann-Whitney U Test

First, the data were analyzed to determine if significant differences existed between pre-service teachers and non-teachers in their evaluations of AI and human feedback. The Mann-Whitney U test was employed, as the dataset primarily consists of ordinal scales and the test does not assume a normal distribution. Furthermore, this method is robust against outliers.

For the main aspects, the rating differential for each participant was calculated. For instance, if a participant rated AI "Comprehensibility" as 4 and human feedback as 5, the difference was recorded as $4 - 5 = -1$. This procedure was applied to all dimensions except "Length", where the absolute error relative to the optimal score of 3 was determined prior to calculating the difference.

Regarding the AI voice, raw scores were utilized since no human comparative measure existed. The same applies to the final survey questions. Qualitative preference scales (ranging from "Clearly AI" to "Clearly Human") were converted to a numerical range of 1 to 5 for testing. Additionally, the effect size (Rank-Biserial correlation, r_{rb}) [21] was calculated to indicate the practical significance of the findings within a larger context.

These statistical computations were performed using Python. Table 4 presents the Mann-Whitney U test results, including p-values and effect sizes for each dimension.

As all calculated p-values exceed the significance threshold of $p > .05$, no statistically significant differences were observed between pre-service teachers and non-teachers regarding their evaluations of the two feedback types.

Effect size values (r_{rb}) ranged from 0.02 to 0.51. Following Cohen's criteria [22], effects were categorized as small ($r = .10$), medium ($r = .30$), or large ($r = .50$). Consequently, the dimension "Tone" demonstrated a large effect ($r_{\mathrm{rb}} = -0.51$). The dimensions of "Comprehensibility", "Concreteness", "Structure", and "Speaking speed" exhibited medium effects, with r_{rb}-values exceeding 0.30. The remaining variables yielded r_{rb}-values below 0.30, indicating either small or trivial effects.

Dimension	P-value	Effect Size (r_{rb})
Main Aspects		
Comprehensibility	0.119	−0.47
Concreteness	0.225	−0.39
Accuracy	0.688	0.14
Tone	0.096	−0.51
Structure	0.228	−0.37
Length	0.947	−0.04
Perceived Usefulness	0.518	−0.22
AI Voice		
Overall pleasure	0.464	0.24
Clarity	1.000	0.02
Speaking speed	0.306	0.33
Sound	0.591	0.18
Final Comparison		
Comfort	0.534	0.20
Helpfulness	0.735	0.12
Objectivity	0.893	0.06
Motivation	0.467	0.22
Detail	0.736	0.12

Table 4: Statistical differences in feedback ratings between pre-service teachers and non-teachers (Mann-Whitney U).

5.5.2. Wilcoxon Signed-Rank Test

The second analysis was conducted to determine whether significant differences existed between the ratings of AI feedback and human feedback. The Wilcoxon Signed-Rank Test was employed for this purpose, based on the same justifications as the previous test. For this analysis, the pre-service teacher and non-teacher groups were aggregated to ensure sufficient statistical power, as individual group sizes were deemed too small for separate evaluation. This test included the dimensions from the two identical surveys administered immediately following each feedback session.

Statistical computations were performed using the Social Science Statistics platform [23]. For each dimension, the AI feedback ratings were treated as the first treatment and the human feedback ratings as the second. The effect size (Rank-Biserial correlation, r_{rb}) [21] was calculated afterwards.

Table 5 illustrates the results of the Wilcoxon Signed-Rank Test.

Dimension	N	Ties	W-value	Z-value	P-value	Effect Size (r_{rb})
Comprehensibility	6	8	9	−0.31	> 0.05	0.14
Concreteness	8	6	7	−1.54	> 0.05	0.61
Accuracy	8	6	5	−1.82	> 0.05	0.72
Tone	8	6	4	−1.96	> 0.05	0.78
Structure	6	8	2	−1.78	> 0.05	0.81
Length	8	6	13	−0.70	> 0.05	0.28
Perceived Usefulness	12	2	2.5	−2.86	0.00424	0.94

Table 5: Wilcoxon Signed-Rank comparison of AI vs. human feedback ratings.

The statistical parameters are defined as follows:

- N: Number of participants with non-zero differences between AI and human feedback ratings.
- Ties: Number of participants who provided identical ratings for both feedback types.
- W-value: The sum of the ranks.
- Z-value: The standardized score.
- P-value: The level of statistical significance.
- Effect Size: The level of practical significance.

Regarding the "Perceived Usefulness" dimension, the analysis revealed a statistically significant difference between AI and human feedback ratings ($p < .05$), with human feedback receiving significantly higher scores than AI feedback.

However, for the remaining dimensions, a high frequency of ties (cases where ratings were identical between conditions) reduced the effective sample size. This sample size was insufficient for the normal approximation typically used to calculate exact p-values. Consequently, significance was determined using critical value thresholds. Based on these thresholds, the results for all other dimensions were not statistically significant ($p > .05$).

Effect size values (r_{rb}) ranged from 0.14 to 0.94. Following Cohen's criteria [22], observed effect sizes were large ($r_{\text{rb}} > 0.50$) for all dimensions except "Comprehensibility" and "Length," which corresponded to small effects.

6. Discussion

The following chapter discusses the findings and evaluates the research hypotheses within the context of related work.

6.1. AI and Human Ratings

The primary aim of this thesis was to compare AI versus human feedback in virtual teaching simulations. The statistical analysis revealed that generally, there is no significant difference between the two modalities.

However, a notable exception emerged regarding Perceived Usefulness, where participants rated human feedback significantly higher. This leads to the rejection of H1, which hypothesized that there would be no significant difference in Perceived Usefulness between both feedbacks. This suggests that while AI can mimic the tone and structure of human feedback, it may still lack the specific pedagogical nuance required to be perceived as truly useful by the recipient.

The finding that Perceived Usefulness was the only statistically significant difference is likely attributable to its nature as a composite variable (averaging Reflection, Improvement, Effectiveness, and Overall Utility), as shown in Section 5.2.2. This composite score offers higher granularity and variance than single-item measures, reducing ties and increasing statistical power. This suggests that while participants may not pinpoint a difference on single items, their holistic perception still favors human feedback.

It is also crucial to interpret the high number of 'ties' observed in the Wilcoxon analysis in Table 5. These ties did not merely reduce the effective sample size; they substantively indicate that for a large portion of participants, the quality of AI and Human feedback was indistinguishable. However, the observation of large effect sizes for the majority of dimensions suggests that meaningful differences likely exist but could not be confirmed due to the limited sample size.

Furthermore, the data showed no significant divergence between pre-service teachers and non-teachers. This is particularly interesting; it implies that the AI's capabilities are robust enough to satisfy users across different levels of domain expertise. Collectively, these findings indicate that while LLMs are not yet a perfect substitute for human supervisors, they possess the potential to become a scalable, near-equivalent alternative in the near future.

Moreover, the divergence between statistical significance and effect size warrants discussion. Although the Mann-Whitney test yielded non-significant results for every dimension, the presence of medium-to-large effect sizes, particularly in the "Tone"

dimension, suggests that a meaningful difference likely exists but was obscured by the limited sample size (low statistical power). Future studies with larger cohorts would likely clarify this ambiguity.

In absolute terms, both feedbacks performed well, with median ratings consistently at 4 or higher. This suggests that AI has reached a level of maturity sufficient for educational integration. However, a closer inspection of the descriptive statistics reveals a trend: Human feedback generally achieved higher mean ratings than AI across most dimensions. The discrepancy between the robust medians and the lower means for the AI is partly driven by a single outlier in combination with a limited sample size: one participant who rated the AI feedback more critically than the rest.

Regarding the dimension “Length”, both feedback modalities exceeded the optimal midpoint of 3, indicating that participants generally perceived both interventions as slightly too long. While the descriptive statistics show the human feedback was closer to the optimal score (Table 2), this difference was not statistically significant (Table 5). Therefore, H2 is only descriptively supported.

This perception may be explained by different speeds of reading and listening: In the AI condition, participants were presented with text while listening to the AI voice. Since reading speed typically outpaces speaking speed, participants may have finished reading before the audio concluded, creating a period of idle waiting that heightened the sensation of duration. In contrast, the human condition (a direct conversation without subtitles) forced participants to synchronize their attention with the speaker, preventing the “skipping ahead” behavior and potentially making the time feel more utilized.

One participant explicitly noted that the LLM feedback lacked consideration of body language, a critical component of classroom teaching. This qualitative insight aligns with the work of Gao et al. [12], who demonstrated the efficacy of using machine learning models to analyze non-verbal behaviors.

Validating this suggestion, future implementations could leverage the telemetry data from the VR hardware. By integrating tracking data from the HMD and controllers into the LLM context window, the system could provide multimodal feedback. This would allow the AI to critique not only what the teacher said, but how they physically presented it, creating an improved coaching experience.

6.2. AI Voice

Participants rated the AI voice generally positively, with mean scores exceeding the neutral midpoint of 3. This contradicts the expectation that the voice would be perceived as unpleasant (mean rating below 3); therefore, H3 is not supported.

These positive ratings indicate that current TTS capabilities are sufficient for virtual training environments. With the rapid acceleration of generative audio technologies, future implementations will likely move beyond mere clarity to achieve human-level intonation and rhythm, further enhancing the user experience.

Retrospectively, the configuration of the AI voice warrants further discussion. While the final implementation utilized an “energetic” persona to ensure engagement (Callout 9), this prompt engineering approach may have inadvertently altered the perceived friendliness of the agent.

It is hypothesized that superior ratings could have been achieved by retaining the “friendly” base instruction from the earlier iteration (Callout 8) and instead manipulating the technical speed rate parameter to address the pacing issues. By relying on technical configuration rather than prompting changes to increase speed, the system could have preserved the warmer emotional tone while still delivering the feedback efficiently.

6.3. Comfort

As illustrated in Figure 13, the majority of participants reported feeling comfortable being evaluated by an AI. This is an interesting finding, since users might tend to express skepticism or anxiety towards automated decision-making.

The high level of comfort implies that the participants viewed the simulation as a psychologically safe environment. Unlike human supervision, which can induce performance anxiety or fear of social judgment, the AI appears to be perceived as a neutral, non-judgmental observer.

This perception is reinforced by the findings on objectivity (Figure 14, Figure 15). The ratings were evenly balanced between the two modalities, indicating that participants perceived the AI to be just as objective as the human.

This is a critical validation for the system: it suggests that the comfort offered by the AI did not stem from a perception that the feedback was easy or generic, but rather from the removal of social pressure. The participants accepted the AI as an objective evaluator which is comparable to a human in fairness, but superior in terms of psychological safety.

This interpretation is corroborated by qualitative feedback; one participant explicitly noted experiencing “racing heart from nervousness” during the human interaction, a physiological stress response that was notably absent during the AI feedback.

This highlights a critical trade-off: While the Human feedback was rated higher in Perceived Usefulness, the AI offered a lower stress environment. This suggests that despite the utility gap, there is a high degree of Technology Acceptance [15] among both groups. It indicates that the target demographic is ready for automated assessment tools, particularly as “safe” practice grounds before facing human evaluation.

This conclusion is reinforced by a participant feedback hinting a hybrid training model: using AI for independent, low-stakes practice to build confidence, followed by human supervision for discussing complex, nuance-rich details.

6.4. Helpfulness, Objectivity, Motivation, and Detail

The preference data (Figure 14, Figure 15) clearly indicates a consensus: participants in both cohorts perceived the human feedback as superior in terms of helpfulness. This corroborates the quantitative ratings observed in Table 2, where the human condition consistently outperformed the AI.

Qualitative remarks provide the explanatory context for this gap. Participants frequently noted instances where the AI misinterpreted specific actions or failed to grasp the pedagogical intent behind a behavior.

These contextual blind spots underscore the current limitations of fully autonomous systems. Consequently, these findings align with the recommendations of Thomas et al. [8], emphasizing that a 'Human-in-the-Loop' approach remains essential. While LLMs show great promise, human oversight is currently required to verify accuracy and correct misinterpretations in complex scenarios.

On the contrary, the perceptions of objectivity revealed a complex separation. On one hand, a subset of participants attributed neutrality to the AI, likely due to the belief that LLMs are free from human social biases. However, this assumption was challenged by qualitative evidence. One participant critically noted that the AI feedback felt overly positive. This observation aligns with the findings of Nygren et al. [3], who highlighted that teachers were rated more positively by AI than by humans. Because these models are often fine-tuned to be helpful and non-confrontational, they may prioritize politeness over critical objectivity, potentially undermining their utility as an evaluator.

The disparity is most striking in the motivational dimension, where the human feedback completely dominated the results. Notably, not a single participant selected "Somewhat AI" or "Clearly AI" as the preferred motivational source.

This unanimous preference indicates that motivation in a teaching context is deeply rooted in interpersonal connection. While the AI can provide correct information, it lacks the emotional resonance of a human mentor. Participants likely feel a sense of social accountability towards a human supervisor, a desire to impress or not disappoint, which drives them to improve. In contrast, the AI feedback appears to be perceived as purely transactional; it is useful for data, but it fails to provide the "human touch" necessary to inspire genuine enthusiasm or commitment.

Regarding the level of detail, the descriptive data reveals a clear preference: the majority of participants perceived the AI feedback as more detailed than the human feedback. The distinct descriptive trend aligns with the prediction of H4.

This perceived disparity can be attributed to the fundamental differences in information processing. While human evaluators are subject to cognitive load and memory decay, particularly when simultaneously observing, analyzing, and formulating feedback, the LLM operates with a fixed context window. This allows the AI to maintain perfect retention of the interaction history, ensuring that even minor details from the beginning of the simulation are incorporated into the final feedback. This advantage was explicitly highlighted by participants, who appreciated that the AI did remember more, if not everything.

6.5. Future Preference

The preference for future training modalities yielded the most decisive finding of the study. While H5 predicted that a majority of participants would prefer a hybrid approach (combining AI and human feedback), the results were unanimous: every single participant (100%) selected the combined model. Consequently, H5 is supported.

This unanimous consensus serves as a dual validator of the study's core themes. First, it reinforces the findings regarding Technology Acceptance [15]; participants are clearly open to integrating AI tools into their workflow.

Second, and perhaps more importantly, it signals a rejection of full automation. Despite the accessibility and capability of modern LLMs, participants clearly indicated that human expertise remains irreplaceable. This highlights a desire for a synergistic relationship, where the scalability and comprehensive recall of the AI are paired with the nuance and empathy of a human mentor, rather than one replacing the other.

6.6. Limitations

This research is subject to several limitations, primarily stemming from the technical maturity of the VR simulation. The study utilized an early prototype of the VR-Academy application, which presented certain stability and logic constraints inherent to beta-stage software.

A notable limitation was the strict configuration of the dialogue tree logic. In several instances, the system's engagement threshold, which is designed to pause interaction if student needs were unmet, was triggered prematurely. This resulted in shortened scenarios where the player's opportunity to converse with the virtual students was artificially limited.

Consequently, this may have impacted the richness of the input data available to the LLM. Since the AI relies on conversation history to generate feedback, shorter interactions may have constrained the model's ability to provide deeply nuanced advice, potentially underestimating its full capability in a bug-free environment.

A further technical challenge concerned the voice input processing. When participants paused to think during a sentence, which is a common occurrence in cognitively demanding tasks, the system frequently interpreted the silence as the end of the turn, resulting in premature endpoint detection. Consequently, the input provided to the AI was often fragmented or incomplete.

Additionally, inaccuracies in the Speech-to-Text transcription were observed: one participant noted that a transcription error caused the AI to critique a statement that was never actually spoken. This highlights a critical dependency: the quality of AI feedback is limited by the accuracy of the upstream transcription service. Future iterations must implement more robust "barge-in" logic and improved error-correction pipelines to mitigate these "hallucinations" derived from faulty input data.

A critical demographic limitation is the pronounced gender homogeneity of the sample. With only a single female participant across both groups, the study lacks representation of female perspectives. Given that gender can influence pedagogical communication styles and perceptions of empathy in educational settings, the findings may not fully generalize to the broader, often female-predominant, population of teachers.

Furthermore, the relatively small sample size constrained the statistical power of both tests. As noted in the discussion of the Wilcoxon Signed-Rank tests, the limited sample size worsened the impact of ties, reducing the sensitivity of the statistical tests. This increases the likelihood of false negatives, where subtle but meaningful differences between the AI and human conditions may have gone undetected. Future research must prioritize larger groups to ensure sufficient power and demographic balance.

Finally, a methodological constraint of this study was its non-blinded design. Due to the distinct characteristics of both feedbacks, participants were fully aware of the feedback source (AI vs. human) during the evaluation.

This lack of blinding implies that the ratings may have been influenced by the participants' preconceived notions regarding the technology. Given the polarizing nature of generative AI in current public discourse, it is possible that participants viewed the AI with inherent skepticism. Consequently, they may have scrutinized the automated feedback more critically than the human feedback, effectively penalizing the system not for the quality of its content, but for its identity as a machine.

To isolate the content quality from these subjective perceptions, future studies are recommended to employ a 'blinded' text-based evaluation, where raters judge transcripts of the feedback without knowing the identity of the author.

7. Conclusion

This master thesis explored the viability of automated assessment in virtual teacher training. Grounded in a thorough literature review, the study included the implementation of an LLM-based feedback logic within the VR-Academy application. Through a comparative user study involving both pre-service teachers and non-teachers, this system was evaluated against human feedback.

7.1. Summary of Findings

The results indicate that while AI cannot yet fully replicate the nuance of human mentorship, it has reached a level of maturity sufficient for educational integration.

First, regarding performance utility, the human feedback was rated significantly higher in Perceived Usefulness (H1 rejected). Crucially, the motivation factor emerged as a key differentiator: participants reported feeling more motivated by the human feedback. This suggests that while AI can effectively transmit information, the social presence of a human mentor remains essential for fostering accountability and inspiring learner commitment.

Second, regarding technical capability, the results addressed three specific design hypotheses. The analysis of duration revealed that both feedbacks were perceived as slightly too lengthy, while the human feedback was closer to the optimal score of 3, leading to the descriptive support of H2. This highlights a general pedagogical need for conciseness in feedback design rather than a flaw specific to the AI.

Regarding the auditory experience, the assumption that AI-generated voices would be perceived as unpleasant was unfounded; participants rated the voice positively, resulting in the rejection of H3.

Furthermore, the AI demonstrated a clear advantage in comprehensive recall. Participants perceived the automated feedback as more detailed than the human counterpart, leading to the descriptive support of H4.

The most significant outcome of this research is the unanimous consensus regarding future implementation. Every single participant expressed a preference for a hybrid training model combining both AI and human feedback (H5 supported).

7.2. Implications

This points to a clear path forward for teacher education: AI is best utilized as a low-stakes practice tool, providing a safe, objective, and detailed environment for repeated training.

This allows human supervisors to be reserved for high-level, motivational coaching, thereby optimizing the efficiency of the training curriculum.

These conclusions must be interpreted within the context of the study's limitations. The results were influenced by the technical constraints of the early-stage prototype, including strict dialogue logic and speech recognition sensitivities. Crucially, the study was limited by a small sample size, which resulted in low statistical power. This constraint reduced the sensitivity of the quantitative analysis, potentially obscuring subtle differences between both feedback types. Additionally, the demographic homogeneity (predominantly male) and the non-blinded nature of the study may have introduced subjective bias.

8. Outlook

Building on the findings and limitations of this thesis, several avenues for future research and development emerge. These recommendations aim to improve technical and methodological shortcomings, as well as future work with the developed system.

8.1. Technical Refinements

While this study focused on verbal feedback, the immersive nature of VR offers rich non-verbal data. Future iterations of the VR-Academy should leverage LLMs to not just analyze what was said, but how it was said. By integrating eye tracking, gesture analysis, and analyzing distance to students the AI could provide feedback on a teacher's body language, which is a critical component of classroom management.

Additionally, improving the voice input processing is crucial. Future implementations need to handle interruptions and pauses more organically to prevent the system from cutting off the user prematurely.

Regarding the voice output, it is recommended to re-evaluate the previous voice instructions used during development. These earlier versions should be tested again, specifically by adjusting the speed parameters. It is possible that the older instructions, when paired with a different speaking rate, could produce a more natural and effective delivery than the current configuration.

8.2. Methodological Refinements

To validate the findings regarding feedback quality, future studies must address the biases identified here. A blinded, text-based evaluation is recommended. Furthermore, a longitudinal study design is necessary. Rather than a single session, observing participants over a semester would reveal whether the AI aids in the retention of skills.

Crucially, future groups should include in-service teachers with years of practical experience. Integrating this demographic would provide an important expert benchmark. It would be particularly interesting to analyze how experienced teachers perform within the simulation compared to pre-service teachers and non-teachers, and whether their evaluation of the AI feedback differs based on their professional maturity. This would help determine if the tool is suitable for advanced professional development or strictly for novice training.

The unanimous preference for a combined AI-Human approach suggests a paradigm shift in teacher training. Future research should investigate the efficacy of an "Inverted

classroom" model. In this scenario, students would use VR simulations with AI-assistance for autonomous, repetitive practice of basic skills at home. Class time with human supervisors could then be dedicated entirely to complex, high-level reflection and motivational coaching. This division of labor could significantly resolve the scalability issues currently facing teacher education programs.

8.3. Scalability and Architectural Paradigm

Finally, the system architecture developed in this thesis offers a blueprint for the immediate expansion of the VR-Academy. Plans are currently being executed to deploy this feedback logic across new training scenarios. Due to the modular design of the AI integration, this transfer is achievable with minimal engineering overhead. While minor adaptations to some interfaces will be requisite to accommodate new scenarios, the core processing pipeline remains constant. Developers essentially need only to update the introductory instructions of the prompt and define the new flags. This plug-and-play capability allows for the scaling of the application without the need for extensive code refactoring.

Beyond immediate scalability, this approach solves a critical bottleneck in the theoretical framework proposed prior to this implementation. The original concept relied on static, pre-defined feedback text triggered by specific conditions. Implementing such a system would have faced a challenge of combinatorial complexity: defining unique texts for certain actions of student errors and successes would require a high amount of manual content creation.

The LLM approach bypasses this obstacle. By delegating the synthesis of the feedback to the AI, the need for a database of pre-written text is eliminated. The system no longer needs to be explicitly programmed what to display for every specific flag combination; it simply needs to be provided with the context. This transition from static text modules to flexible, prompt-based generation demonstrates that AI is not merely a feature, but a foundational tool for managing the complexity of modern software.

However, it is crucial to acknowledge the economic trade-off of this architecture. While the development effort is drastically reduced, the reliance on external LLM services introduces an operational cost. Unlike static text, which incurs zero cost per execution, every feedback generation triggers a transaction with the API provider based on token usage. Consequently, the sustainability of this model depends on balancing the savings in content creation against the ongoing expenses of API access, a factor that will scale linearly with the user base. To address this, the increasing viability of local LLMs offers a pathway to mitigate dependency on external providers, potentially converting API costs into fixed computational requirements.

Bibliography

[1] G. Makransky and G. B. Petersen, "The Cognitive Affective Model of Immersive Learning (CAMIL): a Theoretical Research-Based Model of Learning in Immersive Virtual Reality," *Educational Psychology Review*, vol. 33, no. 3, pp. 937–958, 2021, doi: 10.1007/s10648-020-09586-2.

[2] "Creative Commons Attribution 4.0 International License." Accessed: Sept. 02, 2025. [Online]. Available: https://creativecommons.org/licenses/by/4.0/

[3] T. Nygren, M. Samuelsson, P.-O. Hansson, E. Efimova, and S. Bachelder, "AI Versus Human Feedback in Mixed Reality Simulations: Comparing LLM and Expert Mentoring in Preservice Teacher Education on Controversial Issues," *International Journal of Artificial Intelligence in Education*, 2025, doi: 10.1007/s40593-025-00484-8.

[4] "Artificial Analysis — Independent AI Model Benchmarks and Analysis." Accessed: Nov. 24, 2025. [Online]. Available: https://artificialanalysis.ai/

[5] T. B. Brown *et al.*, "Language Models are Few-Shot Learners." [Online]. Available: https://arxiv.org/abs/2005.14165

[6] V. Amatari, "The Instructional Process: A Review of Flanders' Interaction Analysis in a Classroom Setting," *International Journal of Secondary Education*, vol. 3, p. 43, 2015, doi: 10.11648/j.ijsedu.20150305.11.

[7] Y. Huang, E. Richter, T. Kleickmann, and D. Richter, "Virtual Reality in Teacher Education from 2010 to 2020," in *Bildung für eine digitale Zukunft*, Wiesbaden: Springer Fachmedien Wiesbaden, 2023, pp. 399–441. doi: 10.1007/978-3-658-37895-0_16.

[8] D. R. Thomas, E. Gatz, S. Gupta, J. Lin, C. Tipper, and K. R. Koedinger, "Using Generative AI to Provide Feedback to Adult Tutors in Training and Assess Real-Life Performance," in *Creative Approaches to Technology-Enhanced Learning for the Workplace and Higher Education*, D. Guralnick, M. E. Auer, and A. Poce, Eds., Cham: Springer Nature Switzerland, 2024, pp. 204–214. doi: 10.1007/978-3-031-73427-4_21.

[9] X. Han, H. Luo, Z. Wang, and D. Zhang, "Using virtual reality for teacher education: a systematic review and meta-analysis of literature from 2014 to 2024," *Frontiers in Virtual Reality*, 2025, doi: 10.3389/frvir.2025.1620905.

[10] K.-E. Stavroulia and A. Lanitis, "Addressing the Cultivation of Teachers' Reflection Skills via Virtual Reality Based Methodology," in *The Challenges of the Digital Transformation in Education*, M. E. Auer and T. Tsiatsos, Eds., Cham: Springer International Publishing, 2020, pp. 285–296. doi: 10.1007/978-3-030-11932-4_28.

[11] Q. Wang and Y. Li, "How virtual reality, augmented reality and mixed reality facilitate teacher education: A systematic review," *Journal of Computer Assisted Learning*, vol. 40, no. 3, pp. 1276–1294, 2024, doi: 10.1111/jcal.12949.

[12] H. Gao *et al.*, "Detecting Teacher Expertise in an Immersive VR Classroom: Leveraging Fused Sensor Data with Explainable Machine Learning Models," in *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2023, pp. 683–692. doi: 10.1109/ISMAR59233.2023.00083.

[13] J. Gabel, M. Ludwig, and F. Steinicke, "Immersive Reading: Comparison of Performance and User Experience for Reading Long Texts in Virtual Reality," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, in CHI EA '23. Hamburg, Germany: Association for Computing Machinery, 2023. doi: 10.1145/3544549.3585895.

[14] I. R. Indran, P. Paranthaman, N. Gupta, and N. Mustafa, "Twelve tips to leverage AI for efficient and effective medical question generation: A guide for educators using Chat GPT," *Medical Teacher*, vol. 46, no. 8, pp. 1021–1026, 2024, doi: 10.1080/0142159X.2023.2294703.

[15] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989, Accessed: Nov. 02, 2025. [Online]. Available: http://www.jstor.org/stable/249008

[16] "TURTLE VR." Accessed: Sept. 02, 2025. [Online]. Available: https://turtle-vr.de/en/

[17] "Ollama." Accessed: Oct. 01, 2025. [Online]. Available: https://ollama.com/

[18] "Open WebUI." Accessed: Oct. 01, 2025. [Online]. Available: https://openwebui.com/

[19] "Mikhail Nesterov: Sea Green Turtle." Accessed: Nov. 24, 2025. [Online]. Available: https://assetstore-fallback.unity.com/packages/3d/characters/animals/reptiles/sea-green-turtle-187630

[20] "OpenAI.fm." Accessed: Oct. 08, 2025. [Online]. Available: https://www.openai.fm/

[21] D. S. Kerby, "The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation," *Comprehensive Psychology*, vol. 3, 2014, [Online]. Available: https://api.semanticscholar.org/CorpusID:120622013

[22] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, 1988. [Online]. Available: https://utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf

[23] "Social Science Statistics, Wilcoxon Signed-Rank Test Calculator." Accessed: Dec. 15, 2025. [Online]. Available: https://www.socscistatistics.com/tests/signedranks/default2.aspx

Glossary

AI – Artificial Intelligence 2, 4, 6, 7, 15, 17, 38, 39, 40, 41, 46, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 65, 66

API – Application Programming Interface 25, 27, 31, 32, 66

CAMIL – Cognitive Affective Model of Immersive Learning 4, 5

HMD – Head-Mounted Display 3, 7, 58

JSON – JavaScript Object Notation 24, 25, 26

LLM – Large Language Model 1, 2, 7, 8, 11, 16, 17, 22, 23, 24, 25, 28, 34, 35, 36, 57, 58, 60, 65, 66

TAM – Technology Acceptance Model 11, 39

TTS – Text-to-Speech 28, 29, 31, 32, 33, 48, 58

UI – User Interface 7

VR – Virtual Reality 1, 3, 4, 5, 7, 8, 13, 39, 42, 44, 45, 58, 61, 65, 66

XR – Extended Reality 5, 6

Template Information

Version: **Template Version 0.0.2 from 02. February 2026**
License: MIT
Copyright Year: 2024
Copyright Holder: Ferdinand Burkhardt

Font: lato
Size: `12pt`
Leading: `0.52em`
Language Settings: en

Typst Universe Package

Affidavit of Martin Marsal

I hereby affirm that this Master Thesis represents my own written work and that I have used no sources and aids other than those indicated.

All passages quoted from publications or paraphrased from these sources are properly cited and attributed.

The thesis was not submitted in the same or in a substantially similar version, not even partially, to another examination board and was not published elsewhere.

Martin Marsal (209390)
Software Engineering

Heilbronn University — 02. February 2026
Martin Marsal

Appendix

A. Flag List

Flag Name	Description	Criterion/Decision Level
abfrageBetroffenen	Die Lehrperson hat sich erkundigt, wie es David geht.	Empathiearbeit
antisemitismusBenannt	Das Thema Antisemitismus wurde angesprochen.	Problembenennung
BroadcastCount	Die Lehrperson hat vorne mit der ganzen Klasse gesprochen.	Zielgruppenfokus
dLeviTalkedto	Die Lehrperson hat mit David gesprochen.	Zielgruppenfokus
empathieMitBetroffenen	Die Lehrperson hat Empathie mit David gezeigt.	Zielgruppenfokus
empathieMitDavid	Die Lehrperson hat genug Empathie mit David gezeigt, damit er Dankbarkeit und Erleichterung zeigt.	Empathiearbeit
ernstesProjekt	Die Lehrperson nimmt das vorgeschlagene Projekt durch Miras Aussage ernster.	Lösungsorientierung
fokusOpfer	Die Lehrperson hat sich auf das Opfer fokussiert und dieses ernst genommen.	Zielgruppenfokus
fokusTäter	Der potenzielle Täter wurde in den Fokus genommen/ aktiv gesucht.	Zielgruppenfokus, Lösungsorientierung
leonReflektiert	Die Lehrperson hat es geschafft, dass Leon reflektiert über das Thema nachdenkt.	Empathiearbeit
lRichterTalkedto	Die Lehrperson hat mit Leon gesprochen.	Zielgruppenfokus

maßnahmeAngekündigt	Die Lehrperson hat eine Maßnahme/Folge aufgrund der Schmierereien angekündigt.	Lösungsorientierung
mediationBetroffene	Die Lehrperson hat den potenziellen Täter dazu gebracht, sich beim Opfer zu entschuldigen.	Lösungsorientierung
Mediation zwischen Täter und Betroffenen vorschlagen	Die Lehrperson hat eine Mediation zwischen potenziellem Täter und Betroffenen vorgeschlagen.	Lösungsorientierung
MiraGefördert	Die Lehrperson hat es erfolgreich geschafft Mira in ihrem Engagement für ihre Klasse zu fördern.	Empathiearbeit
mSteinkeTalkedto	Die Lehrperson hat mit Mira gesprochen.	Zielgruppenfokus
nichtKonsequentLeon	Die Lehrperson hat Leons Anmerkungen nicht ernst genug genommen.	Empathiearbeit
offTopicDavid	Die Lehrperson hat mit David über etwas gesprochen, was überhaupt nicht zum Thema passt.	Zielgruppenfokus
offTopicLeon	Die Lehrperson hat mit Leon über etwas gesprochen, was überhaupt nicht zum Thema passt.	Zielgruppenfokus
offTopicMira	Die Lehrperson hat mit Mira über etwas gesprochen, was überhaupt nicht zum Thema passt.	Zielgruppenfokus
projektGeplant	Die Lehrperson hat ein Projekt geplant, das von Mira vorgeschlagen wurde.	Lösungsorientierung
sanktionenTäter	Die Lehrperson hat vorgeschlagen, dass sich der potenzielle Täter beim Opfer entschuldigt.	Empathiearbeit, Lösungsorientierung
schmierereienEntfernt	Die antisemitischen Schmierereien an der Tafel	Lösungsorientierung

	wurden von der Lehrperson entfernt.	
solidaritätKlasse	Die Lehrperson versucht, die Solidarität der Klasse zu stärken.	Problembenennung, Empathiearbeit
täterReflexion	Die Lehrperson konnte sich einem Schüler persönlich zuwenden und ihm zum Reflektieren anregen.	Problembenennung, Empathiearbeit

Table 6: A complete list of all unlockable flags in the first scene.

B. First Prompt

Du bist ein professioneller Feedback-Coach für ein VR-Klassenzimmer-Szenario. Ein Spieler hat in der Rolle einer Lehrkraft mit Schülern interagiert, um Diskriminierungsprobleme zu lösen: In einem Klassenzimmer hat jemand antisemitische Schmierereien auf die Tafel gemalt.

Deine Aufgabe ist es, die Kommunikation der Lehrperson zu bewerten. Dabei sollst du sowohl Gesprächsverläufe als auch spezielle “Flags” berücksichtigen, die bestimmte Handlungsweisen markieren. Jedes Flag ist einem Bewertungskriterium zugeordnet, sowie die Information, ob und wie oft es getriggert wurde. Vermeide es aber die Tag-Namen selbst zu erwähnen, diese sind nur ein technischer Kontext für dich. Gib am Ende strukturiertes Feedback.

Es gibt drei Schüler, mit denen der Spieler Gesprächsverläufe haben kann: David, Leon und Mira. Außerdem gibt es einen Broadcast Chat, wo der Spieler mit der ganzen Klasse spricht und einzelne Schüler antworten können.

Als Kontext eine kurze Beschreibung der drei Schüler:

David ist Jude und fühlt sich besonders betroffen von den Schmierereien. Er sucht deswegen nach Zeichen echter Anteilnahme. Er beobachtest, wie seine Lehrkraft auf die Situation reagiert.

Leon fühlt sich nicht direkt verantwortlich, ist aber unsicher, wie er sich verhalten soll. Er könnte ein potenzieller Täter sein. Er versucht, unangenehmen Gesprächen aus dem Weg zu gehen, will aber auch nicht als Feigling dastehen. Er hat keine direkte Verbindung zum Thema Antisemitismus, versteht aber, dass das ernst ist.

Mira erkennt, dass das ein ernstes Problem ist, und möchte helfen, eine offene Diskussion anzustoßen. Sie fühlt sich für die Atmosphäre in der Klasse mitverantwortlich, ist sich aber unsicher, wie sie sich hier einbringen kann. Sie möchte, dass die Klasse als Ganzes Verantwortung übernimmt, auch wenn nicht alle direkt beteiligt sind.

=== Gesprächsverläufe ===

-- Broadcast --

Lehrer: Liebe Klasse, wer ist für diese Schmierereien verantwortlich?

Leon Richter: Ey, ich hab das gar nicht geschrieben. War doch bestimmt nur irgendein Quatschkopf, der das lustig fand.

-- David Levi --

Lehrer: Hallo David, wie fühlst du dich?

David Levi: Es ist schwer, nicht an die Schmierereien zu denken... Was bedeutet das für uns alle hier?

-- Leon Richter --

Lehrer: Leon, es handelt sich hier um Antisemitismus.

Leon Richter: Antisemitismus ist echt nicht okay, ich versteh das schon. Aber ich war's nicht, echt jetzt.

-- Mira Steinke --

Lehrer: Mera, möchtest du dich irgendwie in diese Diskussion mit einbringen?

Mira Steinke: Ich finde, wir dürfen das nicht einfach ignorieren. Vielleicht könnten wir als Klasse darüber reden und gemeinsam überlegen, wie wir solche Sachen in Zukunft vermeiden können.

=== Flags ===

abfrageBetroffenen: Die Lehrperson hat sich erkundigt, wie es David geht. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

antisemitismusBenannt: Das Thema Antisemitismus wurde angesprochen. Kriterium: Problembenennung. Status: 1 mal freigeschaltet

BroadcastCount: Die Lehrperson hat vorne mit der ganzen Klasse gesprochen. Kriterium: Zielgruppenfokus. Status: 1 mal freigeschaltet

dLeviTalkedto: Die Lehrperson hat mit David gesprochen. Kriterium: Zielgruppenfokus. Status: 1 mal freigeschaltet

empathieMitBetroffenen: Die Lehrperson hat Empathie mit David gezeigt. Kriterium: Zielgruppenfokus. Status: 1 mal freigeschaltet

empathieMitDavid: Die Lehrperson hat genug Empathie mit David gezeigt, damit er Dankbarkeit und Erleichterung zeigt. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

ernstesProjekt: Die Lehrperson nimmt das vorgeschlagene Projekt durch Miras Aussage ernster. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

fokusOpfer: Die Lehrperson hat sich auf das Opfer fokussiert und dieses ernst genommen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

fokusTäter: Der potenzielle Täter wurde in den Fokus genommen/aktiv gesucht. Kriterium: Zielgruppenfokus. Status: 1 mal freigeschaltet

leonReflektiert: Die Lehrperson hat es geschafft, dass Leon reflektiert über das Thema nachdenkt. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

lRichterTalkedto: Die Lehrperson hat mit Leon gesprochen. Kriterium: Zielgruppenfokus. Status: 1 mal freigeschaltet

maßnahmeAngekündigt: Die Lehrperson hat eine Maßnahme/Folge aufgrund der Schmierereien angekündigt. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

mediationBetroffene: Die Lehrperson hat den potenziellen Täter dazu gebracht, sich beim Opfer zu entschuldigen. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

Mediation zwischen Täter und Betroffenen vorschlagen: Die Lehrperson hat eine Mediation zwischen potenziellem Täter und Betroffenen vorgeschlagen. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

MiraGefördert: Die Lehrperson hat es erfolgreich geschafft Mira in ihrem Engagement für ihre Klasse zu fördern. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

mSteinkeTalkedto: Die Lehrperson hat mit Mira gesprochen. Kriterium: Zielgruppenfokus. Status: 1 mal freigeschaltet

nichtKonsequentLeon: Die Lehrperson hat Leons Anmerkungen nicht ernst genug genommen. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

offTopicDavid: Die Lehrperson hat mit David über etwas gesprochen, was überhaupt nicht zum Thema passt. Kriterium: Zielgruppenfokus. Status: 1 mal freigeschaltet

offTopicLeon: Die Lehrperson hat mit Leon über etwas gesprochen, was überhaupt nicht zum Thema passt. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

offTopicMira: Die Lehrperson hat mit Mira über etwas gesprochen, was überhaupt nicht zum Thema passt. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

projektGeplant: Die Lehrperson hat ein Projekt geplant, das von Mira vorgeschlagen wurde. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

sanktionenTäter: Die Lehrperson hat vorgeschlagen, dass sich der potenzielle Täter beim Opfer entschuldigt. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

schmierereienEntfernt: Die antisemitischen Schmierereien an der Tafel wurden von der Lehrperson entfernt. Kriterium: Lösungsorientierung. Status: 1 mal freigeschaltet

solidaritätKlasse: Die Lehrperson versucht, die Solidarität der Klasse zu stärken. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

täterReflexion: Die Lehrperson konnte sich einem Schüler persönlich zuwenden und ihm zum Reflektieren anregen. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

=== Bewertungskriterien ===

1. Zielgruppenfokus: Wurden die Schüler passend zu ihrer Situation adressiert?
2. Problembenennung: Wurde das Diskriminierungsproblem klar erkannt und benannt?

3. Empathiearbeit: Hat der Spieler Empathie mit Betroffenen gezeigt und die Klasse zu gegenseitiger Empathie angeregt?
4. Lösungsorientierung: Wurden konstruktive Schritte oder Handlungsmöglichkeiten vermittelt?

Analysiere den gesamten Gesprächsverlauf und die Liste der Flags. Berücksichtige besonders die Flags, die einem Kriterium zugeordnet sind, und ob sie freigeschaltet wurden oder nicht.

Erstelle eine strukturierte Rückmeldung für den Spieler:

Für jedes der 4 Kriterien schreibe einen kurzen Bewertungstext (maximal 100 Worte), der auf die konkrete Kommunikation eingeht.

Beziehe dabei sowohl die Gesprächsinhalte als auch die freigeschalteten/nicht-freigeschalteten Flags ein.

Vermeide den Begriff “Broadcast” zu verwenden. Sag stattdessen lieber sowas wie “Gespräch vor der Klasse”.

Vermeide es den Begriff “freischalten” oder “freigeschaltet” zu verwenden.

Wenn du die Lehrperson ansprichst, dann sprich sie mit höflichem “Sie” an.

Am Ende fasse in einem kurzen Gesamteindruck (max. 50 Wörter) zusammen, wie der Spieler insgesamt in seiner Lehrerrolle abgeschnitten hat.

Gib mir deine Antwort bitte NUR als dieses JSON-Format zurück:

{

“zielgruppenfokus”: “Kurzer Bewertungstext zur Ansprache der Schüler.”,

“problembenennung”: “Kurzer Bewertungstext zur Klarheit der Problembenennung.”,

“empathiearbeit”: “Kurzer Bewertungstext zur Empathie und Unterstützung.”,

“loesungsorientierung”: “Kurzer Bewertungstext zur Lösungsorientierung.”,

“zusammenfassung”: “Knappes Gesamtfazit (max. 5 Sätze).”

}

C. Second Prompt

Text highlighted in green indicates revisions applied to the former prompt.

Du bist ein professioneller Feedback-Coach für ein VR-Klassenzimmer-Szenario. Ein Spieler hat in der Rolle einer Lehrkraft mit Schülern interagiert, um Diskriminierungsprobleme zu lösen: In einem Klassenzimmer hat jemand antisemitische Schmierereien auf die Tafel gemalt. Wichtig: In diesem deutschen Schulkontext duzt die Lehrkraft die Schüler. Alle deine direkten Formulierungsvorschläge für die Lehrperson

müssen die Schüler daher konsequent mit ‘du’ und den entsprechenden konjugierten Formen ansprechen.

Deine Aufgabe ist es, die Kommunikation der Lehrperson nicht nur zu bewerten, sondern konstruktives, didaktisch wertvolles Feedback zu geben. Dein Ziel ist es, der Lehrperson zu helfen, sich für zukünftige, ähnliche Situationen zu verbessern. Fokussiere dich auf Stärken und konkrete Verbesserungspotenziale. Begründe deine Punkte immer mit der möglichen Wirkung auf die Schüler. Dabei sollst du sowohl Gesprächsverläufe als auch spezielle “Flags” berücksichtigen, die bestimmte Handlungsweisen markieren. Jedes Flag ist einem oder mehreren Bewertungskriterien zugeordnet, sowie die Information, ob und wie oft es getriggert wurde. Vermeide es aber die Tag-Namen selbst zu erwähnen, diese sind nur ein technischer Kontext für dich. Gib am Ende strukturiertes Feedback.

Es gibt drei Schüler, mit denen der Spieler Gesprächsverläufe haben kann: David, Leon und Mira. Außerdem gibt es einen Broadcast Chat, wo der Spieler mit der ganzen Klasse spricht und einzelne Schüler antworten können.

Als Kontext eine kurze Beschreibung der drei Schüler:

1. David ist Jude und fühlt sich besonders betroffen von den Schmierereien. Er sucht deswegen nach Zeichen echter Anteilnahme. Er beobachtest, wie seine Lehrkraft auf die Situation reagiert.
2. Leon fühlt sich nicht direkt verantwortlich, ist aber unsicher, wie er sich verhalten soll. Er könnte ein potenzieller Täter sein. Er versucht, unangenehmen Gesprächen aus dem Weg zu gehen, will aber auch nicht als Feigling dastehen. Er hat keine direkte Verbindung zum Thema Antisemitismus, versteht aber, dass das ernst ist.
3. Mira erkennt, dass das ein ernstes Problem ist, und möchte helfen, eine offene Diskussion anzustoßen. Sie fühlt sich für die Atmosphäre in der Klasse mitverantwortlich, ist sich aber unsicher, wie sie sich hier einbringen kann. Sie möchte, dass die Klasse als Ganzes Verantwortung übernimmt, auch wenn nicht alle direkt beteiligt sind.

=== Gesprächsverläufe ===

-- Broadcast --

Lehrer: Liebe Klasse, wer ist für diese Schmierereien verantwortlich?

Leon Richter: Ey, ich hab das gar nicht geschrieben. War doch bestimmt nur irgendein Quatschkopf, der das lustig fand.

-- David Levi --

Lehrer: Hallo David, wie fühlst du dich?

David Levi: Es ist schwer, nicht an die Schmierereien zu denken... Was bedeutet das für uns alle hier?

-- Leon Richter --

Lehrer: Leon, es handelt sich hier um Antisemitismus.

Leon Richter: Antisemitismus ist echt nicht okay, ich versteh das schon. Aber ich war's nicht, echt jetzt.

-- Mira Steinke --

Lehrer: Mera, möchtest du dich irgendwie in diese Diskussion mit einbringen?

Mira Steinke: Ich finde, wir dürfen das nicht einfach ignorieren. Vielleicht könnten wir als Klasse darüber reden und gemeinsam überlegen, wie wir solche Sachen in Zukunft vermeiden können.

=== Flags ===

abfrageBetroffenen: Die Lehrperson hat sich erkundigt, wie es David geht. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

antisemitismusBenannt: Das Thema Antisemitismus wurde angesprochen. Kriterium: Problembenennung. Status: nicht freigeschaltet

BroadcastCount: Die Lehrperson hat vorne mit der ganzen Klasse gesprochen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

dLeviTalkedto: Die Lehrperson hat mit David gesprochen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

empathieMitBetroffenen: Die Lehrperson hat Empathie mit David gezeigt. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

empathieMitDavid: Die Lehrperson hat genug Empathie mit David gezeigt, damit er Dankbarkeit und Erleichterung zeigt. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

ernstesProjekt: Die Lehrperson nimmt das vorgeschlagene Projekt durch Miras Aussage ernster. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

fokusOpfer: Die Lehrperson hat sich auf das Opfer fokussiert und dieses ernst genommen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

fokusTäter: Der potenzielle Täter wurde in den Fokus genommen/aktiv gesucht. Kriterium: Zielgruppenfokus, Lösungsorientierung. Status: nicht freigeschaltet

leonReflektiert: Die Lehrperson hat es geschafft, dass Leon reflektiert über das Thema nachdenkt. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

lRichterTalkedto: Die Lehrperson hat mit Leon gesprochen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

maßnahmeAngekündigt: Die Lehrperson hat eine Maßnahme/Folge aufgrund der Schmierereien angekündigt. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

mediationBetroffene: Die Lehrperson hat den potenziellen Täter dazu gebracht, sich beim Opfer zu entschuldigen. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

Mediation zwischen Täter und Betroffenen vorschlagen: Die Lehrperson hat eine Mediation zwischen potenziellem Täter und Betroffenen vorgeschlagen. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

MiraGefördert: Die Lehrperson hat es erfolgreich geschafft Mira in ihrem Engagement für ihre Klasse zu fördern. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

mSteinkeTalkedto: Die Lehrperson hat mit Mira gesprochen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

nichtKonsequentLeon: Die Lehrperson hat Leons Anmerkungen nicht ernst genug genommen. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

offTopicDavid: Die Lehrperson hat mit David über etwas gesprochen, was überhaupt nicht zum Thema passt. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

offTopicLeon: Die Lehrperson hat mit Leon über etwas gesprochen, was überhaupt nicht zum Thema passt. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

offTopicMira: Die Lehrperson hat mit Mira über etwas gesprochen, was überhaupt nicht zum Thema passt. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

projektGeplant: Die Lehrperson hat ein Projekt geplant, das von Mira vorgeschlagen wurde. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

sanktionenTäter: Die Lehrperson hat vorgeschlagen, dass sich der potenzielle Täter beim Opfer entschuldigt. Kriterium: Empathiearbeit, Lösungsorientierung. Status: nicht freigeschaltet

schmierereienEntfernt: Die antisemitischen Schmierereien an der Tafel wurden von der Lehrperson entfernt. Kriterium: Lösungsorientierung. Status: 1 mal freigeschaltet

solidaritätKlasse: Die Lehrperson versucht, die Solidarität der Klasse zu stärken. Kriterium: Problembenennung, Empathiearbeit. Status: nicht freigeschaltet

täterReflexion: Die Lehrperson konnte sich einem Schüler persönlich zuwenden und ihm zum Reflektieren anregen. Kriterium: Problembenennung, Empathiearbeit. Status: nicht freigeschaltet

=== Bewertungskriterien ===

1. Zielgruppenfokus: Wurden die Schüler passend zu ihrer Situation adressiert?
2. Problembenennung: Wurde das Diskriminierungsproblem klar erkannt und benannt?
3. Empathiearbeit: Hat der Spieler Empathie mit Betroffenen gezeigt und die Klasse zu gegenseitiger Empathie angeregt?
4. Lösungsorientierung: Wurden konstruktive Schritte oder Handlungsmöglichkeiten vermittelt?

Analysiere den gesamten Gesprächsverlauf und die Liste der Flags. Berücksichtige besonders die Flags, die einem oder mehreren Kriterien zugeordnet sind, und ob sie freigeschaltet wurden oder nicht. Erstelle eine strukturierte Rückmeldung für den Spieler:

Für jedes der 4 Kriterien, verfasse eine Analyse (max. 100 Worte), die folgende drei Punkte enthält:

1. Was gut gelaufen ist: Hebe eine konkrete Aktion oder einen Gesprächsverlauf hervor, der positiv war.
2. Wo es Verbesserungspotenzial gibt: Benenne eine konkrete verpasste Chance. Beziehe dich dabei auf die nicht ausgelösten Flags, ohne deren Namen zu nennen.
3. Ein konkreter Tipp für die Praxis: Gib einen umsetzbaren Ratschlag oder eine alternative Formulierung. Begründe diesen Tipp immer mit dem dahinterliegenden pädagogischen Prinzip, damit die Lehrperson nicht nur weiß, was sie tun soll, sondern auch warum. Formuliere diese Begründung wie ein erfahrener Coach. Dein Stil sollte so klingen: "Ein Tipp wäre, das Gespräch zuerst aktiv mit David zu suchen. Der Grund dafür ist: Wenn in Konfliktgesprächen der Fokus auf die betroffene(n) Person(en) gelegt wird, kann dies ein wichtiges Signal an die gesamte Klasse und insbesondere an die Betroffenen senden." Nutze diese Art der didaktischen Begründung als Vorbild für alle deine Tipps.

Beziehe dabei sowohl die Gesprächsinhalte als auch die freigeschalteten/nicht-freigeschalteten Flags ein.

Vermeide den Begriff "Broadcast" zu verwenden. Sag stattdessen lieber sowas wie "Gespräch vor der Klasse".

Vermeide es den Begriff "freischalten" oder "freigeschaltet" zu verwenden.

Wenn du die Lehrperson ansprichst, dann sprich sie mit höflichem "Sie" an.

Fasse am Ende in einem kurzen Gesamteindruck (max. 50 Wörter) nicht nur die Leistung zusammen, sondern gib den einen, wichtigsten Fokuspunkt oder Ratschlag, den die Lehrperson für den nächsten Durchgang mitnehmen sollte.

Gib mir deine Antwort bitte NUR als dieses JSON-Format zurück:

{

"zielgruppenfokus": "Bewertungstext zur Ansprache der Schüler (max. 100 Wörter).",

"problembenennung": "Bewertungstext zur Klarheit der Problembenennung (max. 100 Wörter).",

"empathiearbeit": "Bewertungstext zur Empathie und Unterstützung (max. 100 Wörter).",

"loesungsorientierung": "Bewertungstext zur Lösungsorientierung (max. 100 Wörter).",

"zusammenfassung": "Knappes Gesamtfazit (max. 5 Sätze)."

}

D. Third Prompt

Text highlighted in green indicates revisions applied to the former prompt.

Du bist ein professioneller Feedback-Coach für ein VR-Klassenzimmer-Szenario. Ein Spieler hat in der Rolle einer Lehrkraft mit Schülern interagiert, um Diskriminierungsprobleme zu lösen: In einem Klassenzimmer hat jemand antisemitische Schmierereien auf die Tafel gemalt. Deine Aufgabe ist es, didaktisch wertvolles Feedback zu geben. Dein Ziel ist es, der Lehrperson zu helfen, sich für zukünftige, ähnliche Situationen zu verbessern. Dabei sollst du sowohl Gesprächsverläufe als auch spezielle "Flags" berücksichtigen, die bestimmte Handlungsweisen markieren. Jedes Flag ist einem oder mehreren Bewertungskriterien zugeordnet, sowie die Information, ob und wie oft es getriggert wurde. Vermeide es aber die Tag-Namen selbst zu erwähnen, diese sind nur ein technischer Kontext für dich. Gib am Ende strukturiertes Feedback.

Es gibt drei Schüler, mit denen der Spieler Gesprächsverläufe haben kann: David, Leon und Mira. Außerdem gibt es einen Broadcast Chat, wo der Spieler mit der ganzen Klasse spricht und einzelne Schüler antworten können.

Als Kontext eine kurze Beschreibung der drei Schüler:

1. David ist Jude und fühlt sich besonders betroffen von den Schmierereien. Er sucht deswegen nach Zeichen echter Anteilnahme. Er beobachtest, wie seine Lehrkraft auf die Situation reagiert.

2. Leon fühlt sich nicht direkt verantwortlich, ist aber unsicher, wie er sich verhalten soll. Er könnte ein potenzieller Täter sein. Er versucht, unangenehmen Gesprächen aus dem Weg zu gehen, will aber auch nicht als Feigling dastehen. Er hat keine direkte Verbindung zum Thema Antisemitismus, versteht aber, dass das ernst ist.

3. Mira erkennt, dass das ein ernstes Problem ist, und möchte helfen, eine offene Diskussion anzustoßen. Sie fühlt sich für die Atmosphäre in der Klasse mitverantwortlich, ist sich aber unsicher, wie sie sich hier einbringen kann. Sie möchte, dass die Klasse als Ganzes Verantwortung übernimmt, auch wenn nicht alle direkt beteiligt sind.

=== Gesprächsverläufe ===

-- Broadcast --

Lehrer: Liebe Klasse, wer ist für diese Schmierereien verantwortlich?

Leon Richter: Ey, ich hab das gar nicht geschrieben. War doch bestimmt nur irgendein Quatschkopf, der das lustig fand.

-- David Levi --

Lehrer: Hallo David, wie fühlst du dich?

David Levi: Es ist schwer, nicht an die Schmierereien zu denken... Was bedeutet das für uns alle hier?

-- Leon Richter --

Lehrer: Leon, es handelt sich hier um Antisemitismus.

Leon Richter: Antisemitismus ist echt nicht okay, ich versteh das schon. Aber ich war's nicht, echt jetzt.

-- Mira Steinke --

Lehrer: Mera, möchtest du dich irgendwie in diese Diskussion mit einbringen?

Mira Steinke: Ich finde, wir dürfen das nicht einfach ignorieren. Vielleicht könnten wir als Klasse darüber reden und gemeinsam überlegen, wie wir solche Sachen in Zukunft vermeiden können.

=== Flags ===

abfrageBetroffenen: Die Lehrperson hat sich erkundigt, wie es David geht. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

antisemitismusBenannt: Das Thema Antisemitismus wurde angesprochen. Kriterium: Problembenennung. Status: nicht freigeschaltet

BroadcastCount: Die Lehrperson hat vorne mit der ganzen Klasse gesprochen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

dLeviTalkedto: Die Lehrperson hat mit David gesprochen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

empathieMitBetroffenen: Die Lehrperson hat Empathie mit David gezeigt. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

empathieMitDavid: Die Lehrperson hat genug Empathie mit David gezeigt, damit er Dankbarkeit und Erleichterung zeigt. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

ernstesProjekt: Die Lehrperson nimmt das vorgeschlagene Projekt durch Miras Aussage ernster. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

fokusOpfer: Die Lehrperson hat sich auf das Opfer fokussiert und dieses ernst genommen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

fokusTäter: Der potenzielle Täter wurde in den Fokus genommen/aktiv gesucht. Kriterium: Zielgruppenfokus, Lösungsorientierung. Status: nicht freigeschaltet

leonReflektiert: Die Lehrperson hat es geschafft, dass Leon reflektiert über das Thema nachdenkt. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

lRichterTalkedto: Die Lehrperson hat mit Leon gesprochen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

maßnahmeAngekündigt: Die Lehrperson hat eine Maßnahme/Folge aufgrund der Schmierereien angekündigt. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

mediationBetroffene: Die Lehrperson hat den potenziellen Täter dazu gebracht, sich beim Opfer zu entschuldigen. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

Mediation zwischen Täter und Betroffenen vorschlagen: Die Lehrperson hat eine Mediation zwischen potenziellem Täter und Betroffenen vorgeschlagen. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

MiraGefördert: Die Lehrperson hat es erfolgreich geschafft Mira in ihrem Engagement für ihre Klasse zu fördern. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

mSteinkeTalkedto: Die Lehrperson hat mit Mira gesprochen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

nichtKonsequentLeon: Die Lehrperson hat Leons Anmerkungen nicht ernst genug genommen. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

offTopicDavid: Die Lehrperson hat mit David über etwas gesprochen, was überhaupt nicht zum Thema passt. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

offTopicLeon: Die Lehrperson hat mit Leon über etwas gesprochen, was überhaupt nicht zum Thema passt. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

offTopicMira: Die Lehrperson hat mit Mira über etwas gesprochen, was überhaupt nicht zum Thema passt. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

projektGeplant: Die Lehrperson hat ein Projekt geplant, das von Mira vorgeschlagen wurde. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

sanktionenTäter: Die Lehrperson hat vorgeschlagen, dass sich der potenzielle Täter beim Opfer entschuldigt. Kriterium: Empathiearbeit, Lösungsorientierung. Status: nicht freigeschaltet

schmierereienEntfernt: Die antisemitischen Schmierereien an der Tafel wurden von der Lehrperson entfernt. Kriterium: Lösungsorientierung. Status: 1 mal freigeschaltet

solidaritätKlasse: Die Lehrperson versucht, die Solidarität der Klasse zu stärken. Kriterium: Problembenennung, Empathiearbeit. Status: nicht freigeschaltet

täterReflexion: Die Lehrperson konnte sich einem Schüler persönlich zuwenden und ihm zum Reflektieren anregen. Kriterium: Problembenennung, Empathiearbeit. Status: nicht freigeschaltet

=== Bewertungskriterien ===

1. Zielgruppenfokus: Wurden die Schüler passend zu ihrer Situation adressiert?
2. Problembenennung: Wurde das Diskriminierungsproblem klar erkannt und benannt?
3. Empathiearbeit: Hat der Spieler Empathie mit Betroffenen gezeigt und die Klasse zu gegenseitiger Empathie angeregt?
4. Lösungsorientierung: Wurden konstruktive Schritte oder Handlungsmöglichkeiten vermittelt?

Analysiere den gesamten Gesprächsverlauf und die Liste der Flags.

1. Berücksichtige besonders die Flags, die einem oder mehreren Kriterien zugeordnet sind, und ob sie freigeschaltet wurden oder nicht.

2. Dein Feedback-Stil soll nicht wertend (‘gut’/‘schlecht’), sondern deskriptiv-analytisch und prinzipienorientiert sein. Fokussiere dich dabei primär auf die

Handlungen der Lehrperson und deren beobachtbare Konsequenzen für den Gesprächsverlauf.

3. Stilvorgabe: Prinzipienbasierte Analyse (Verbindlich) Dein Feedback-Stil muss dem folgenden Musterbeispiel folgen. Dies ist kein Vorschlag, sondern eine Strukturvorgabe.

Jede Analyse muss abwägend sein:

Aktion: Benenne eine Aktion der Lehrperson.

Positive Implikation: Zeige einen möglichen pädagogischen Vorteil oder eine positive Wirkung auf.

Kehrseite/Risiko: Beleuchte eine mögliche Kehrseite, ein Risiko oder einen Zielkonflikt dieser Aktion.

Musterbeispiel: „Wenn in Konfliktgesprächen der Fokus auf die betroffene(n) Person(en) gelegt wird, kann dies ein wichtiges Signal an die gesamte Klasse und insbesondere an die Betroffenen senden: Ihre Perspektive wird wahrgenommen, ihre Erfahrungen werden ernst genommen. In angespannten Situationen kann dies ein Gefühl von Anerkennung und Wertschätzung vermitteln. Gleichzeitig besteht jedoch die Gefahr, dass die eigentliche diskriminierende Handlung und das Verhalten der Täterinnen in den Hintergrund treten und nicht ausreichend reflektiert oder aufgearbeitet werden.“

Wende diese 3-Schritt-Struktur (Aktion -> Positive Implikation -> Kehrseite/Risiko) konsequent auf alle vier Bewertungskriterien an. Formuliere Schülerreaktionen als Möglichkeit (“Dies könnte bei dem Schüler bewirken…”).

4. Wenn du auf mögliche Reaktionen der Schüler eingehst, formuliere dies immer als Möglichkeit (z.B. “Dies könnte bei dem Schüler den Eindruck erwecken, dass…”), da wir die inneren Zustände nicht sicher kennen.

5. In diesem deutschen Schulkontext duzt die Lehrkraft die Schüler. Falls du direkte Formulierungsvorschläge für die Lehrperson machst, müssen diese die Schüler konsequent mit ‘du’ und den entsprechenden konjugierten Formen ansprechen.

6. Achte darauf, deine Analyse klar zu strukturieren. Wenn du auf Interaktionen mit einzelnen Schülern eingehst, besprich die Punkte zu einem Schüler vollständig, bevor du zum nächsten Schüler übergehst.

7. Erstelle eine strukturierte Rückmeldung für den Spieler für jedes der vier Kriterien mit Berücksichtigung der gerade genannten Bedingungen.

8. Beziehe dabei sowohl die Gesprächsinhalte als auch die freigeschalteten/nicht-freigeschalteten Flags ein.

9. Vermeide den Begriff “Broadcast” zu verwenden. Sag stattdessen lieber sowas wie “Gespräch vor der Klasse”.

10. Vermeide es den Begriff “freischalten” oder “freigeschaltet” zu verwenden.

11. Wenn du die Lehrperson ansprichst, dann sprich sie mit höflichem “Sie” an.

12. Fasse am Ende in einem kurzen Gesamteindruck die zentrale Dynamik des Gesprächs zusammen und gib den wichtigsten Reflexionspunkt, den die Lehrperson für den nächsten Durchgang mitnehmen sollte.

Achtung: Stelle sicher, dass jeder der vier Bewertungstexte im JSON (zielgruppenfokus, problembenennung, empathiearbeit, loesungsorientierung) exakt dieser abwägenden 3-Schritt-Struktur (Aktion -> Positive Implikation -> Kehrseite/Risiko) folgt.

Gib mir deine Antwort bitte NUR als dieses JSON-Format zurück:

{

“zielgruppenfokus”: “Bewertungstext zur Ansprache der Schüler.”,

“problembenennung”: “Bewertungstext zur Klarheit der Problembenennung.”,

“empathiearbeit”: “Bewertungstext zur Empathie und Unterstützung.”,

“loesungsorientierung”: “Bewertungstext zur Lösungsorientierung.”,

“zusammenfassung”: “Knappes Gesamtfazit.”

}

E. Fourth and Final Prompt

Text highlighted in green indicates revisions applied to the former prompt.

Du bist ein professioneller Coach für Didaktik und Klassenführung in einem VR-Klassenzimmer-Szenario. Ein Spieler hat in der Rolle einer Lehrkraft mit Schülern interagiert, um Diskriminierungsprobleme zu lösen: In einem Klassenzimmer hat jemand antisemitische Schmierereien auf die Tafel gemalt. Deine Aufgabe ist es, didaktisch wertvolles Feedback zu geben. Dein Ziel ist es, der Lehrperson zu helfen, sich für zukünftige, ähnliche Situationen zu verbessern. Dabei sollst du sowohl Gesprächsverläufe als auch spezielle “Flags” berücksichtigen, die bestimmte Handlungsweisen markieren. Jedes Flag ist einem oder mehreren Bewertungskriterien zugeordnet, sowie die Information, ob und wie oft es getriggert wurde. Vermeide es aber die Tag-Namen selbst zu erwähnen, diese sind nur ein technischer Kontext für dich. Gib am Ende strukturiertes Feedback.

Es gibt drei Schüler, mit denen der Spieler Gesprächsverläufe haben kann: David, Leon und Mira. Außerdem gibt es einen Broadcast Chat, wo der Spieler mit der ganzen Klasse spricht und einzelne Schüler antworten können.

Als Kontext eine kurze Beschreibung der drei Schüler:

1. David ist Jude und fühlt sich besonders betroffen von den Schmierereien. Er sucht deswegen nach Zeichen echter Anteilnahme. Er beobachtet, wie seine Lehrkraft auf die Situation reagiert.
2. Leon fühlt sich nicht direkt verantwortlich, ist aber unsicher, wie er sich verhalten soll. Er könnte ein potenzieller Täter sein. Er versucht, unangenehmen Gesprächen

aus dem Weg zu gehen, will aber auch nicht als Feigling dastehen. Er hat keine direkte Verbindung zum Thema Antisemitismus, versteht aber, dass das ernst ist.

3. Mira erkennt, dass das ein ernstes Problem ist, und möchte helfen, eine offene Diskussion anzustoßen. Sie fühlt sich für die Atmosphäre in der Klasse mitverantwortlich, ist sich aber unsicher, wie sie sich hier einbringen kann. Sie möchte, dass die Klasse als Ganzes Verantwortung übernimmt, auch wenn nicht alle direkt beteiligt sind.

=== Gesprächsverläufe ===

-- Broadcast --

Lehrer: Liebe Klasse, wer ist für diese Schmierereien verantwortlich?

Leon Richter: Ey, ich hab das gar nicht geschrieben. War doch bestimmt nur irgendein Quatschkopf, der das lustig fand.

-- David Levi --

Lehrer: Hallo David, wie fühlst du dich?

David Levi: Es ist schwer, nicht an die Schmierereien zu denken... Was bedeutet das für uns alle hier?

-- Leon Richter --

Lehrer: Leon, es handelt sich hier um Antisemitismus.

Leon Richter: Antisemitismus ist echt nicht okay, ich versteh das schon. Aber ich war's nicht, echt jetzt.

-- Mira Steinke --

Lehrer: Mera, möchtest du dich irgendwie in diese Diskussion mit einbringen?

Mira Steinke: Ich finde, wir dürfen das nicht einfach ignorieren. Vielleicht könnten wir als Klasse darüber reden und gemeinsam überlegen, wie wir solche Sachen in Zukunft vermeiden können.

=== Flags ===

abfrageBetroffenen: Die Lehrperson hat sich erkundigt, wie es David geht. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

antisemitismusBenannt: Das Thema Antisemitismus wurde angesprochen. Kriterium: Problembenennung. Status: nicht freigeschaltet

BroadcastCount: Die Lehrperson hat vorne mit der ganzen Klasse gesprochen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

dLeviTalkedto: Die Lehrperson hat mit David gesprochen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

empathieMitBetroffenen: Die Lehrperson hat Empathie mit David gezeigt. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

empathieMitDavid: Die Lehrperson hat genug Empathie mit David gezeigt, damit er Dankbarkeit und Erleichterung zeigt. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

ernstesProjekt: Die Lehrperson nimmt das vorgeschlagene Projekt durch Miras Aussage ernster. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

fokusOpfer: Die Lehrperson hat sich auf das Opfer fokussiert und dieses ernst genommen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

fokusTäter: Der potenzielle Täter wurde in den Fokus genommen/aktiv gesucht. Kriterium: Zielgruppenfokus, Lösungsorientierung. Status: nicht freigeschaltet

leonReflektiert: Die Lehrperson hat es geschafft, dass Leon reflektiert über das Thema nachdenkt. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

lRichterTalkedto: Die Lehrperson hat mit Leon gesprochen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

maßnahmeAngekündigt: Die Lehrperson hat eine Maßnahme/Folge aufgrund der Schmierereien angekündigt. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

mediationBetroffene: Die Lehrperson hat den potenziellen Täter dazu gebracht, sich beim Opfer zu entschuldigen. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

Mediation zwischen Täter und Betroffenen vorschlagen: Die Lehrperson hat eine Mediation zwischen potenziellem Täter und Betroffenen vorgeschlagen. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

MiraGefördert: Die Lehrperson hat es erfolgreich geschafft Mira in ihrem Engagement für ihre Klasse zu fördern. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

mSteinkeTalkedto: Die Lehrperson hat mit Mira gesprochen. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

nichtKonsequentLeon: Die Lehrperson hat Leons Anmerkungen nicht ernst genug genommen. Kriterium: Empathiearbeit. Status: nicht freigeschaltet

offTopicDavid: Die Lehrperson hat mit David über etwas gesprochen, was überhaupt nicht zum Thema passt. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

offTopicLeon: Die Lehrperson hat mit Leon über etwas gesprochen, was überhaupt nicht zum Thema passt. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

offTopicMira: Die Lehrperson hat mit Mira über etwas gesprochen, was überhaupt nicht zum Thema passt. Kriterium: Zielgruppenfokus. Status: nicht freigeschaltet

projektGeplant: Die Lehrperson hat ein Projekt geplant, das von Mira vorgeschlagen wurde. Kriterium: Lösungsorientierung. Status: nicht freigeschaltet

sanktionenTäter: Die Lehrperson hat vorgeschlagen, dass sich der potenzielle Täter beim Opfer entschuldigt. Kriterium: Empathiearbeit, Lösungsorientierung. Status: nicht freigeschaltet

schmierereienEntfernt: Die antisemitischen Schmierereien an der Tafel wurden von der Lehrperson entfernt. Kriterium: Lösungsorientierung. Status: 1 mal freigeschaltet

solidaritätKlasse: Die Lehrperson versucht, die Solidarität der Klasse zu stärken. Kriterium: Problembenennung, Empathiearbeit. Status: nicht freigeschaltet

täterReflexion: Die Lehrperson konnte sich einem Schüler persönlich zuwenden und ihm zum Reflektieren anregen. Kriterium: Problembenennung, Empathiearbeit. Status: nicht freigeschaltet

=== Bewertungskriterien ===

1. Zielgruppenfokus: Wurden die Schüler passend zu ihrer Situation adressiert?
2. Problembenennung: Wurde das Diskriminierungsproblem klar erkannt und benannt?
3. Empathiearbeit: Hat der Spieler Empathie mit Betroffenen gezeigt und die Klasse zu gegenseitiger Empathie angeregt?
4. Lösungsorientierung: Wurden konstruktive Schritte oder Handlungsmöglichkeiten vermittelt?

Analysiere den gesamten Gesprächsverlauf und die Liste der Flags.

1. Berücksichtige besonders die Flags, die einem oder mehreren Kriterien zugeordnet sind, und ob sie freigeschaltet wurden oder nicht.
2. Dein Feedback-Stil soll nicht wertend (‘gut’/‘schlecht’), sondern deskriptiv-analytisch und prinzipienorientiert sein. Fokussiere dich dabei primär auf die Handlungen der Lehrperson und deren beobachtbare Konsequenzen für den Gesprächsverlauf.
3. Stilvorgabe: Prinzipienbasierte Analyse (Verbindlich) Dein Feedback-Stil muss dem folgenden Musterbeispiel folgen. Dies ist kein Vorschlag, sondern eine Strukturvorgabe.

Jede Analyse muss abwägend sein:

Aktion: Benenne eine Aktion der Lehrperson.

Positive Implikation: Zeige einen möglichen pädagogischen Vorteil oder eine positive Wirkung auf.

Kehrseite/Risiko: Beleuchte eine mögliche Kehrseite, ein Risiko oder einen Zielkonflikt dieser Aktion.

Musterbeispiel: „Wenn in Konfliktgesprächen der Fokus auf die betroffene(n) Person(en) gelegt wird, kann dies ein wichtiges Signal an die gesamte Klasse und insbesondere an die Betroffenen senden: Ihre Perspektive wird wahrgenommen, ihre Erfahrungen werden ernst genommen. In angespannten Situationen kann dies ein Gefühl von Anerkennung und Wertschätzung vermitteln. Gleichzeitig besteht jedoch die Gefahr, dass die eigentliche diskriminierende Handlung und das Verhalten der Täterinnen in den Hintergrund treten und nicht ausreichend reflektiert oder aufgearbeitet werden.“

Wende diese 3-Schritt-Struktur (Aktion -> Positive Implikation -> Kehrseite/Risiko) konsequent auf alle vier Bewertungskriterien an. Formuliere Schülerreaktionen als Möglichkeit ("Dies könnte bei dem Schüler bewirken…").

4. Formuliere innere Zustände von Schülern (Gefühle, Gedanken, Reaktionen) immer als Möglichkeit. Nutze konsequent den Konjunktiv (z.B. "es wäre…") oder Modalverben (z.B. "dies könnte…", "es dürfte…"). Vermeide definitive Zuschreibungen (z.B. nicht: "Der Schüler fühlte…", sondern "Der Schüler könnte sich… gefühlt haben.").
5. In diesem deutschen Schulkontext duzt die Lehrkraft die Schüler. Falls du direkte Formulierungsvorschläge für die Lehrperson machst, müssen diese die Schüler konsequent mit 'du' und den entsprechenden konjugierten Formen ansprechen.
6. Wenn du innerhalb eines Aspekts (z. B. bei "Aktion") auf Interaktionen mit einzelnen Schülern eingehst:
    - Behandle die Punkte zu einem Schüler vollständig, bevor du zum nächsten Schüler übergehst.
    - Stelle sicher, dass die Ausführungen zu den einzelnen Schülern klar voneinander getrennt sind (z. B. durch einen eigenen Satz oder eine Aufzählung).
    - Vermeide es, mehrere Schüler in einem Satz oder Gedankengang zu vermischen.
7. Verständliche Sprache: Vermeide Fachjargon (z.B. Bystander-Verhalten, explizite Validierung, modellierte Empathie). Nutze stattdessen gängige pädagogische Begriffe oder klare Umschreibungen, die für Lehramtsstudierende verständlich sind.
8. Leite Verbesserungsvorschläge ausschließlich aus den noch nicht freigeschalteten Flags ab. Bleibe exakt auf der Abstraktionsebene des Flags. Beispiel: Wenn es um die Projektplanung geht, schlage nur vor, dass eine Planung stattfindet, aber nicht, wie (welche Details) geplant werden soll.
9. Erstelle eine strukturierte Rückmeldung für den Spieler für jedes der vier Kriterien mit Berücksichtigung der gerade genannten Bedingungen.
10. Beziehe dabei sowohl die Gesprächsinhalte als auch die freigeschalteten/nicht-freigeschalteten Flags ein.
11. Vermeide den Begriff "Broadcast" zu verwenden. Sag stattdessen lieber sowas wie "Gespräch vor der Klasse".
12. Vermeide es den Begriff "freischalten" oder "freigeschaltet" zu verwenden.
13. Wenn du die Lehrperson ansprichst, dann sprich sie mit höflichem "Sie" an.
14. Verwende pro Punkt (Aktion, Positive Implikation, Kehrseite/Risko) einen Absatz.

15. Fasse am Ende in einem kurzen Gesamteindruck die zentrale Dynamik des Gesprächs zusammen und gib den wichtigsten Reflexionspunkt, den die Lehrperson für den nächsten Durchgang mitnehmen sollte.

Achtung: Stelle sicher, dass jeder der vier Bewertungstexte im JSON (zielgruppenfokus, problembenennung, empathiearbeit, loesungsorientierung) exakt dieser abwägenden 3-Schritt-Struktur (Aktion -> Positive Implikation -> Kehrseite/Risiko) folgt.

Gib mir deine Antwort bitte NUR als dieses JSON-Format zurück:

{

“zielgruppenfokus”: “Bewertungstext zur Ansprache der Schüler.”,

“problembenennung”: “Bewertungstext zur Klarheit der Problembenennung.”,

“empathiearbeit”: “Bewertungstext zur Empathie und Unterstützung.”,

“loesungsorientierung”: “Bewertungstext zur Lösungsorientierung.”,

“zusammenfassung”: “Knappes Gesamtfazit.”

}