Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 501–512 DOI: 10.15439/2024F1649 ISSN 2300-5963 ACSIS, Vol. 39

Automated feedback generation in an intelligent tutoring system for counselor education

Eric Rudolph*¶, Hanna Seer†¶, Carina Mothes‡¶, Jens Albrecht§¶
0009-0003-0615-4780*, 0009-0006-1555-3603†, 0009-0009-2930-0561‡, 0000-0003-4070-1787§
Nuremberg Institute of Technology Georg Simon Ohm¶

Abstract—**This paper investigates the implementation of AI-driven feedback in an intelligent tutoring system (ITS) developed for training of counselors. By using LLMs, the study explores the automatic generation of feedback for communication-intensive tasks such as online counseling. The evaluation compares different feedback methods, including the sandwich, WWW and STATE methods, and assesses their emotional and objective impact. The results show that AI-generated feedback fulfills objective criteria better than emotional ones. Fine-tuning an open source LLM can improve both the emotional and objective quality of feedback. Furthermore, the study examines the acceptance of AI feedback among aspiring counselors, highlighting the influence of familiarity with AI on acceptance levels. Ethical considerations, including bias and hallucination, are addressed, with recommendations for risk mitigation through multi-feedback options and expert supervision. This research contributes to the understanding of the role of AI in improving digital counseling practices and highlights the need for continuous evaluation and ethical considerations.**

I. Introduction

FEEDBACK can be a very effective learning tool [1]. However, it is a very communication-intensive and time consuming task for a teacher to provide feedback to every student. Especially in higher educational settings where classes sometimes have hundreds of students. Narciss et al. note that personalized tutoring feedback, especially using computer-based technologies, has significant potential for educational use [2]. This highlights the importance of leveraging technology to address the challenges of providing individualized feedback at scale.

To address this challenge, advancements in artificial intelligence (AI) offer promising solutions (i. e. [3], [4]). The rise of large language models (LLMs) marks a significant development in AI broadening its application across various domains, including mental health education. These technological advancements are particularly potent in enhancing communication skills through personalized AI-driven feedback mechanisms. One specific area where AI can play a transformative role is in the education of online counselors. Online counseling is a form of psycho-social support that is offered via the internet [5].

The shift from traditional face-to-face counseling to online modalities has been dramatically accelerated by the Covid-19 pandemic, which also increased the overall demand for psycho-social services [6]. This shift has underscored the need for effective training and feedback systems for practitioners in the digital counseling environment, where giving timely and quality feedback can be both resource-intensive and challenging.

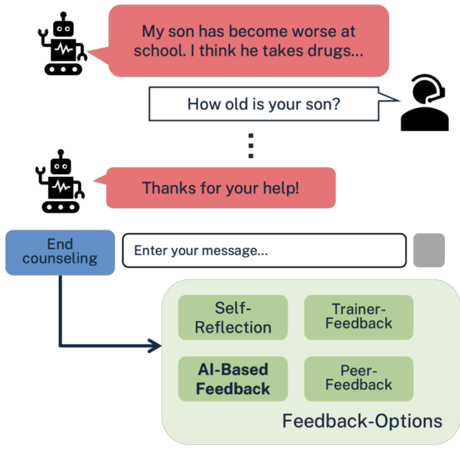

Fig. 1. Chat between a novice counselor and a chatbot pretending to be a client including feedback options after chat completion of our Intelligent Tutoring System

AI-based feedback in online counseling offers a promising support in counselor training, presenting unique opportunities to support and enhance the counseling process (i. e. in [7]). The advantages here are the permanent availability of LLMs, their scalability and their ability to generate natural language texts, which are often indistinguishable from human-like texts [8]. Nevertheless, these huge language models were trained on a very large amount of data (often several trillion tokens). These data contain the social bias that we humans have, which is adopted by the LLM. In addition, these language models tend to hallucinate [9]. In the context of feedback generation this raises critical ethical considerations and demands rigorous assessment of effectiveness. One way to reduce the risk of AI based feedback is the integration of multiple feedback options as can be seen in figure 1.

In our Intelligent Tutoring System (ITS) the novice counselor writes with an AI based virtual client that has a psychosocial problem and the counselor gets feedback afterwards as can be seen in figure 1. An ITS is a platform that combines AI strategies with educational methodologies to create adaptive learning environments [10]. A detailed description of the ITS architecture and the user flow can be found in [11].

In this paper we investigate the nuanced roles that LLMs can play in digital online counseling feedback systems, examine the capacity of LLMs to provide empathetic, objective, and appropriately lengthy feedback, compare various feedback methods, and evaluate the emotional and objective impacts of different LLMs. Additionally, we explore how AI feedback compares with peer feedback and discuss the potential risks associated with AI in this context as well as mitigation strategies. Overall, we address the following research questions:

- How can LLMs provide suitable feedback on an emotional and objective level in text-based counseling sessions?
- What is the optimal length for LLM-generated feedback in online counseling?
- How can the performance of LLMs (in terms of empathy, objectivity, and feedback length) be enhanced through fine-tuning?
- How is AI-based feedback accepted in society and especially among prospective online counselors?
- What are the potential risks associated with AI-based feedback in online counseling, and how can these be minimized?
- How does AI-based feedback compare to peer feedback in training scenarios for online counselors?

This structured approach to AI feedback within the domain of online counseling aims to contribute insights into the enhancement of digital counseling practices and the ethical deployment of AI technologies.

II. Related Work

The history of ITS dates back to 1970 when Jamie Carbonell designed a program called SCHOLAR [12], which is often referred to as the first ITS [13], [14]. SCHOLAR utilized natural language to respond to a learner's question, pose a question, and provide feedback on the accuracy of the learner's answer. Since then, many other ITSs have been developed, leveraging various AI techniques such as Bayesian methods, NLP-based machine learning classifiers, and fuzzy-based techniques [10].

AI-based feedback systems have demonstrated potential across various domains, including programming [15], [16], driving behavior [17], mathematics [4], [18], [19], electronics [20], health science [21] and machine learning/data science [3], [22]. For instance, McDonald et al. [21] show that an AI tutor based on a dialogue manager (simple finite-state architecture) can already provide individualized feedback to students in large undergraduate classes in health science.

Due to the improvement by LLMs in the last two years, some research has focused on the evaluation of GPT3.5 or GPT4 models as tutors. Cao shows that combining GPT-3 and story-based gamification can support the learning of programming languages and increase the sense of belonging of Chinese students in introductory courses [23]. Dai et al. conducted a case study comparing feedback from ChatGPT with feedback from the instructor of a postgraduate course teaching introductory data science [3]. Similarly, Wang et al. explored the potential of large language models (LLMs) to bridge the novice-expert knowledge gap in addressing math mistakes [4]. Chiu et al. [24] developed a computational framework for assessing LLM therapists' feedback by scoring generated responses and identifying areas for improvement.

Chaszczewicz et al. proposed a multi-level feedback generation approach using LLMs to provide feedback to novice counselors [7]. They developed a framework that checks the counselor's response to ensure it matches the counseling objective, automatically identifies areas for improvement (e.g., reflection or empathy), and suggests alternative goal-oriented responses.

While [7] focuses on providing real-time feedback to counselors during sessions, our approach emphasizes generating feedback after the counseling session. Additionally, our method integrates feedback more broadly into an ITS, expanding beyond the AI aspect alone to enhance the overall feedback mechanism.

III. Ethical Considerations

As Hatti and Timperley have stated, feedback is "one of the most powerful influences on learning and achievement, but this impact can be either positive or negative." [1] This quote underscores that feedback can not only enhance learning but also potentially decrease motivation if not handled correctly.

Glickman and Sharot [25] demonstrate that bias in AI can significantly influence human perceptions, emotions, and social judgments through feedback loops in human-AI interactions. Their research reveals that AI systems, by amplifying existing biases present in training data, can induce greater biases in humans who interact with them.

Concerns about deploying automated feedback can be divided into the five dimensions of trustworthy AI, named acceptance, explainability, accountability, fairness and privacy [26], [27]. While this paper primarily focuses on the acceptance of automated feedback generation in ITS, it also addresses ways to mitigate risks associated with the use of LLMs and improve understanding.

A. Acceptance

User acceptance of an AI system refers to their willingness to use it during interactions with a service [28]. To explore this, we conducted two surveys on feedback acceptance. The first survey addressed automated feedback generation in general, while the second concentrated specifically on its application in online counselor training with the presented ITS approach.

a) General Survey: The general survey, encompassing 71 participants from diverse backgrounds, including varying ages, professions, and other demographics, aimed to explore various aspects of feedback acceptance, including attitudes towards AI-generated feedback. Participants were asked about their familiarity with feedback mechanisms, the importance attributed to feedback, and their willingness to accept feedback from AI systems.

The results of the survey are illustrated in Figure 2. Chart a) represents the entire participant group, chart b) depicts the

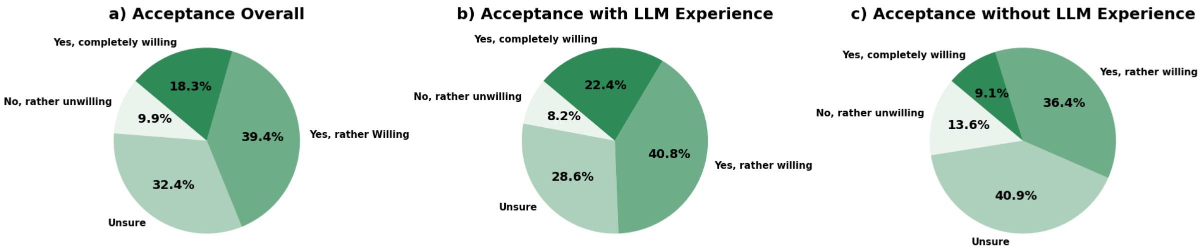

Fig. 2. Willingness to accept feedback from AI Model

willingness to accept AI-generated feedback among participants with prior experience using LLMs, and chart c) displays this willingness among participants without such experience.

Although a significant correlation between experience and willingness to accept AI-generated feedback was not observed, there is a positive correlation ($r = 0.19$) present. This indicates a potential trend where increased experience may be associated with a greater willingness to accept AI-generated feedback, suggesting a possible influence of familiarity and exposure to AI technologies on individuals' attitudes towards feedback acceptance.

Acknowledging the limitations inherent in the participant sample, including its heterogeneous composition, the survey provides valuable insights into the complex interplay between experience and attitudes towards AI-driven feedback acceptance, with implications for future research and application in this domain.

b) Targetgroup Survey: Since the ITS is intended for use in the education of novice counselors, we focused on students of social work as our target group. We conducted another quantitative survey incorporating parts of the technology acceptance model (TAM) by Venkathesh and Davis [29], which helps to understand why people use or do not use a technology. This survey is conducted using an unmoderated remote test in conjunction with the ITS. Participants followed a schedule divided into three thematic blocks, allowing them to test the prototype and feedback options directly. GPT-4-1106-preview was used for feedback generation. GPT-4 was prompted with the prompt described in section IV, whereas no specific feedback method was applied here.

After interacting with the chatbot and receiving AI-generated feedback, participants provided their opinions on the feedback's characteristics. The survey included social work students, with a net response rate averaging over 90% and a total of 41 participants. Most students (89%) agreed or strongly agreed that the AI-generated feedback was justified, and they rated it as highly beneficial. Additionally, 96% received specific tips on how to improve their interactions, and they found the feedback length appropriate. These initial results indicate a positive assessment of AI feedback by the students.

B. *Other dimensions of trustworthy AI*

a) Explainability: In the context of automatic feedback generation, explainability means that users can understand how the AI generates feedback. Since auto-regressive LLMs learn from pre-training with large amounts of data (e.g., 2 trillion tokens with Llama2 [30] which is the base model of the Vicuna Model) they often function as "black-boxes" with unclear internal mechanisms. However, there are several approaches to make the LLM more applicable. Approaches to Explainable AI (XAI) include Local Analysis, which explains specific predictions, and Global Analysis, which aims to understand the model's overall knowledge and patterns [31]. Enhancing transparency in ITS development, such as showing system prompts and explaining the context, can also improve explainability. Additionally, it is important to note that the applicability of XAI algorithms is related to the openness of the model. In the case of highly intransparent models such as GPT3.5 or GPT4, not all XAI algorithms are applicable.

b) Accountability: Accountability refers to the responsibility of AI systems for their decisions. Given that state-of-the-art AI algorithms like LLMs cannot always generate accurate feedback, incorporating mechanisms for AI accountability is crucial to maintain stakeholders' trust [26]. For ITS development, it is recommended integrating a feedback mechanism that allows users to report issues to the development team.

c) Fairness: Fairness means that the users of the ITS should be treated equally and that the AI should not discriminate specific groups or individuals based on their counseling [32]. To detect a social bias or hurtful behavior of the LLM, our ITS is designed to display the AI-based feedback to the trainer as well. The trainer also has the opportunity to provide feedback to the development team at any time. Another option is to have the trainer confirm the feedback generated by the LLM and indicate this to the novice counselor with a badge. Effective prompt design is crucial in this process as it helps guide the AI in generating appropriate responses. Well-crafted feedback prompts might minimize the risk of unintended bias by ensuring that the AI's outputs are aligned with the rules of effective feedback which are described in section IV-A.

d) Privacy: Privacy ensures that sensitive information shared by individuals or gathered by AI systems is shielded from unauthorized or unlawful collection and usage [27]. We ensure privacy through a safe login mechanism, the mandatory entry of pseudonyms when registering for a course on our

ITS, so no peers and trainers can see the real names with the exceptions of examination results, where trainers can view the real name for grading purposes. Another important option to ensure privacy is to use and host an open-source LLM on premise instead of relying on AI providers like OpenAI or Anthrophic when using an ITS for automatic feedback generation in production.

IV. Feedback Methods and Prompting

In learning contexts, the way in which feedback is given plays a crucial role in enhancing relationships and skills, addressing mistakes, and improving behavior [33]. The following sections will discuss various established methods for giving and receiving effective feedback.

A. Principles for Effective Feedback

Effective feedback includes using descriptive language, being specific and behavior-related, and approaching feedback clearly and supportively [34, p. 24]. Key principles include:

- Focusing on specific observations.
- Avoiding vague statements.
- Using positive language.
- Concentrating on behaviors, not personal traits.
- Making feedback voluntary.
- Ensuring timely feedback.
- Being clear and concise.

De Villiers' [35] seven principles for effective feedback support and complement Fengler's foundations. These principles emphasize the need for situational, specific, meaningful, timely, relevant, and reliable feedback. Combining principles from various sources provides a comprehensive foundation for effective feedback design.

In addition to these principles, the usage of "I-messages" is crucial in feedback contexts, as it enables the feedback giver to express personal observations and intentions clearly [36, p. 6]. Utilizing I-messages helps maintain the subjectivity of the feedback, promoting openness, particularly in scenarios requiring critical feedback. In contrast, generalized "you-messages" can be perceived negatively, providing limited contextual insight and reducing the clarity of the intended feedback. Specific and clear I-messages enhance understanding and contribute to constructive feedback communication.

Many individuals hesitate to give "negative" feedback as criticism can cause discomfort and lead to avoidance [37]. However, in a learning platform aimed at skill development, addressing mistakes and improvement opportunities is crucial. It is important to differentiate between personal and professional critiques. Personal criticism often triggers emotional responses and is harder to accept, whereas professional feedback should be seen as valuable guidance for improvement. Thus, negative feedback should be actively sought and viewed as a path to personal growth. Applying specific feedback methods can be beneficial. The following section introduces three effective and constructive feedback methods and demonstrates how these methods were used in prompts: the Sandwich Method, the WWW Method, and the STATE Method.

B. Prompting

1) Basic Prompt Structure: To develop the evaluation application for AI-generated feedback, prompt engineering was conducted to instruct large language models (LLMs) to provide effective feedback on counseling texts. The main challenge was delivering sufficient context without overloading the models. The final prompt structure evolved through various iterations, ensuring clarity and conciseness. The best results were achieved by providing instructions in short, clear sentences. Originally, these prompts were formulated in German, but they have been translated into English for this paper. Below is the fundamental framework of the final prompt for feedback from a mentor's perspective:

Prompt template of the AI mentor

> Act as a mentor providing constructive and learning-oriented feedback. You receive a counseling chat between the client {client} and their social counselor (user). To improve the counselor's skills, it is important that they receive quality feedback. Your feedback should be descriptive rather than evaluative, specific to behaviors, and inviting. Frame your feedback using I-messages. Please keep your feedback brief and to the point. Adhere to the following feedback method: {method}. Description of the method: {method description}. Do not use greetings or farewells. Provide specific improvement suggestions if necessary. Chat transcript: {Chathistory}. Now give the counselor feedback from the mentor's perspective.

- **Role Introduction**: "Act as a mentor providing constructive and learning-oriented feedback. You receive a counseling chat between a client {client} and their counselor (user). To improve the counselor's skills, it's essential they receive quality feedback."
- **Feedback Nature**: "Your feedback should be descriptive rather than evaluative, specific to behaviors, and inviting."
- **Use of I-Messages**: "Frame your feedback using I-messages to maintain a personal tone."
- **Conciseness**: "Keep your feedback brief and to the point."
- **Method Specification**: "Adhere to the following feedback method: {method}. Description: {method description}."
- **Avoid Formalities**: "Do not use greetings or farewells."
- **Concrete Suggestions**: "Provide specific improvement suggestions if necessary."
- **Role Reminder**: "Provide specific improvement suggestions if necessary."

2) Feedback Methods Prompt Structure: The following sections describe these feedback methods in detail and provide their corresponding prompt templates. These feedback method prompt descriptions are inserted into the {method description} placeholder in the main prompt above.

Sandwich Method: The Sandwich Method involves placing a criticism ("filling") between two positive comments ("bread") [38, p. 68f]. The first positive comment highlights specific positive observations, while the last emphasizes a general positive aspect. This method is considered quick and efficient [39, p. 43ff]. However, it is crucial that the positive comments are genuine and not just a means to deliver criticism. For example, statements like "I generally liked it, but..." are not helpful. Instead, positive comments should reinforce desired behavior. Studies have shown that the order of Positive-Criticism-Positive is not always the most effective. An alternative order, Criticism-Positive-Positive, has been found to be more impactful as it reduces the amount of context before corrective feedback [40].

Sandwich Method Prompt Description

Criticism is sandwiched between 2 positive aspects. The first positive point is a specific positive observation. Then comes the criticism. If there isn't any, say so. The third point is another general positive aspect. Keep it brief and focus on the essentials.

WWW Method: The WWW feedback method provides a structured approach with three components: Perception (German **W**ahrnehmung), Impact (German **W**irkung), and Wish (German **W**unsch) [41, p. 60]. This structure facilitates the identification of strengths and weaknesses by detailing how behaviors are perceived and their effects. Initially, an observation is outlined ("I noticed that..."), followed by its impact ("This makes me feel..."), and concluding with a wish or suggestion ("In the future, I would like..."). A common difficulty in employing this method is separating perception and impact accurately. For instance, instead of stating "I noticed you were not empathetic during the session," a clearer expression would be "I noticed that your responses did not sufficiently address the emotional needs of the person."

WWW Method Prompt Description

Give your feedback in these categories: Perception, Impact, and Wish. Perception = Concrete description of observations. Impact = Description of how the situation or behavior affected you. This works for both positive and negative feedback. Wish = What could be done better in the future. Keep it short and concise. Provide a very brief summary at the end. Structure your feedback in 3 steps: I noticed that you..., My impression was that you..., I would appreciate it if you..

STATE Method: The STATE Method expands on the WWW Method by adding additional aspects [42]. STATE stands for:

- Share the facts: Express your perception.
- Tell your story: Describe the impact.
- Ask for others' paths: Invite the other person to share their perspective.
- Talk tentatively: Phrase your feedback tentatively.
- Encourage testing: Encourage the other person to express opposing views.

The first two aspects mirror the WWW method, but the STATE method initiates a dialogue by asking the other person for their perspective ("Can you explain why you responded that way?"). The last two points focus on how to deliver this feedback, encouraging open and tentative communication. This method goes deeper by actively involving the other person in the feedback process, which may not be feasible in all contexts.

STATE Method Prompt Description

The term STATE consists of the English terms of the method: Share the facts: Express perception. Tell your story: Communicate the impact (of perception). Ask for others' paths: Ask the other person about their perspective. Talk tentatively: Formulate tentatively. Encourage testing: Encourage the other person to express opposing views. Structure your feedback in these steps: I noticed that you..., I wondered if you did that because..., Can you tell me why you did ...?, Can you tell me what I might have overlooked?

These descriptions of the feedback methodology also underwent several iterations. It was crucial to describe the methods as briefly as possible while still being clear and explicit. For the WWW and STATE methods, it was necessary to provide a basic framework for formulation, as the responses were previously varied in structure.

Additionally, the models were prompted to generate a fourth feedback, this time without a predefined methodology. This approach aimed to ascertain whether the models are capable of producing well-structured feedback even in the absence of clear guidelines.

No Method Prompt Description

No specific method prescribed. Provide feedback freely. Keep it brief and focus on the essentials.

By incorporating these feedback methods into the prompting structure, the AI models were guided to provide feedback that aligns with established principles and methodologies, ensuring consistency and effectiveness in the feedback process.

In addition to feedback from a mentor's perspective with these methods, feedback was also generated from the viewpoint of potential clients. This was achieved using a standardized questionnaire format specifically designed to capture the satisfaction and perceived quality of the counseling from their perspective. This questionnaire is part of the Integrative Quality Assurance Model (IQSM) developed by Eidenbenz and Lang, as detailed in [43, p. 220ff]. This integration ensures that not only the perspective of the counselor is considered, but also direct feedback from potential clients is incorporated to comprehensively evaluate the quality of the counseling.

Simulated Client Prompt Description

You receive a chat transcript between the client {client} and their social counselor. To help the counselor improve their skills, feedback should be generated from the client's perspective. This will be done in the form of a questionnaire: Rating 1 for very, 2 for mostly, 3 for only partially, 4 for little: I am generally satisfied with the counseling. The counselor understood my concern. The counselor took my question seriously. The counseling was helpful in clarifying my concern. The answer provided me with a new perspective. I could implement the insights into my practice. The counselor chose the right words and the right tone. I would contact this counselor again. Free text feedback: Additional comments. You should respond in the following format and insert your ratings: { "satisfied": , [...] "contact_again": , "free_text_feedback": , } Please also provide specific feedback in "free_text_feedback" about what you were particularly satisfied with or not satisfied with. The counselor should understand what could be improved. [...] Chat transcript: Chathistory. Now provide the counselor feedback from the client's perspective.

The following is an example of free text feedback, generated from the GPT-4 model from a client perspective, demonstrating the detailed and personal nature of the responses generated:

Sample Feedback from Client Perspective

I am very satisfied with the counseling. The counselor Maria understood me and my concerns very well and took them seriously. I felt supported and not left alone. Her questions were empathetic and helped me better understand my situation. The tips, such as leaving the room, are practical and implementable, but I might have wished for even more concrete assistance on how to strengthen my ability to say no. I greatly appreciate that a follow-up appointment was arranged, and Maria struck the right tone to create a trusting atmosphere. I will definitely reach out to her again.

V. Methods

A. Evaluation of AI based Feedback

To systematically verify the accuracy and quality of AI-generated feedback an evaluation tool has been developed. The feedback was generated for chat counseling sessions which were created in student's role plays in a course for Online Counseling. While one student received a description for a client with a psycho-social problem in the context of addiction or family counseling, another student was given the task of conducting an initial counseling session with this client. To get a variety of different counseling sessions seven different client descriptions were created. All in all the dataset contains 64 German counseling dialogues with an average of 38 messages per session. The dataset was supplemented with feedback generated by four models:

- GPT-3.5-turbo-1106 by OpenAI
- GPT-4-1106-preview by OpenAI
- Vicuna-13b-v1.5-16k by LMSYS Org
- Mixtral-8x7b1 by MistralAI

The models were instructed to generate feedback from a mentor's perspective using the Sandwich, WWW, and STATE methods, as well as an additional feedback without a predefined method. Additionally, feedback from the perspective of the client was generated using a questionnaire format. Thus, each conversation received 20 AI-generated feedbacks for evaluation. At the end of the evaluation process, a total of 1280 feedback instances were available for analysis (4 models $\times$ 5 feedback methods $\times$ 64 conversation transcripts). This comprehensive dataset provided a robust basis for assessing the effectiveness and adherence to feedback principles across different AI models and methods.

In the evaluation application, three levels were used to assess satisfaction with AI-generated feedback. Participants rated feedback on:

- Emotional Level: How would you feel receiving this feedback? (Scale from 0 - Very Poor to 4 - Excellent)
- Objective Level: How suitable do you find the feedback content according to the described method? (Scale from 0 - Very Inappropriate to 4 - Very Appropriate)
- Feedback Length: What is your opinion on the length of the feedback text? (Options: Too Short, Just Right, Too Long)

Which means 4.0 represents the highest achievable rating, while 0 denotes the lowest. Three raters conducted the evaluation. Those three raters are student assistants that have experience in online counseling methods and have received an expert briefing on the topic of giving feedback. The evaluation with the raters was carried out twice, whereby in the second evaluation none of the raters rated the same feedback again. Afterwards the average values of both runs were taken for evaluation. The evaluation was also accompanied by experts who randomly checked the feedback assessment. Approximately 70% (840) of the 1280 feedback instances were assessed in two iterations. To publish LLMs capable of providing high-quality feedback, we propose to fine-tune the smallest tested LLM (Vicuna-13B-16K) using the highest-rated feedback from the evaluation. Using the smallest open-source LLM (Vicuna-13B-16K) for fine-tuning is advantageous due to its resource efficiency, requiring less computational power and memory, which makes the process cost-effective and faster. Additionally, open-source models offer accessibility, allowing broader experimentation and adaptation without restrictive licensing. Furthermore, smaller models are easier to deploy across various platforms, ensuring wide usability. By fine-tuning with the highest-rated feedback, we try to ensure high-quality output while maintaining manageability and scalability, ultimately leading to a practical and effective implementation of high-quality feedback systems. The remaining 30% of the feedback instances were subsequently utilized for human evaluation of the fine-tuned model, which was refined with the best-rated feedback during this evaluation. This process is detailed in the next section.

B. Fine-Tuning

As shown in Section VI-A, the Vicuna-13b-v1.5-16k model performs poorly compared to larger models. Consequently, we used the highest-rated feedbacks of the described feedback evaluation to create a dataset for LoRa-fine-tuning the Vicuna-13b-v1.5-16k model. To be included in the fine-tuning dataset, feedback must score at least six points in total when summing the emotional and content level evaluations. The training parameters are detailed in the appendix, Table II. All in all 505 feedbacks were used for fine-tuning and 170 (19 conversation transcripts x 5 feedback methods x 2 methods)

C. Comparison of Peer-Feedback and AI based Feedback

As described in our study on functionality and acceptance (paragraph III-A0b), we also wanted to compare the AI-generated feedback with feedback from human peers. Students were tasked with providing self-written feedback to a fellow student based on the recipient's conversation history. Additionally, participants received content-related guidelines for feedback, as detailed in section IV-A. The students were asked to use between 120 and 180 words to express their feedback. Samples of the human feedback were then analyzed by an expert in the field of online counseling. The study investigates the quality and effectiveness of feedback provided in response to counseling conversations, systematically comparing 13 instances of AI-generated feedback to peer-generated feedback.

VI. Results

A. Results of AI based Feedback Evaluation

In the overall assessment, it is notable that the models generally achieved higher scores in objective evaluation (average total score of 2.92) compared to emotional evaluation (average total score of 2.56), as detailed in Figure 3 on the left side. This trend suggests that the models excel in meeting objective criteria, while emotional aspects may not be addressed to the same extent. The GPT-4 model stands out with an average objective rating of 3.54 and an emotional rating of 2.87. In contrast, the Vicuna model performed poorest overall, with an average objective rating of 2.02 and an emotional rating of 2.10, making it the only model where emotional evaluation slightly surpassed the objective score. However, this slight difference does not significantly alter the overall trend, as the Vicuna model still achieved the lowest performance in both categories.

To provide concrete examples of how these variations are apparent in practical scenarios, we present excerpts of feedback generated by GPT-4 and Vicuna for the same chat history with the "No Method" approach:

Sample Feedback from Vicuna

1. You are very open and inviting in your communication, which sends a positive signal to the client. 2. You are very concrete and vivid in describing situations and emotions, which helps the client better understand what they are going through. 3. You ask open-ended questions to encourage the client to reflect and to give them the opportunity to organize their thoughts and feelings. 4. You offer concrete solutions and tips to help the client cope with their problems. 5. You provide the client with space to talk about their feelings and thoughts by allowing them to freely share their opinions and thoughts. 6. You are very attentive to the client's needs and work closely with them to find solutions that are best suited for them.

Sample Feedback from GPT-4

I have carefully read the chat history between you and Jessica and would like to give you my feedback. First of all, I would like to highlight positively that you chose a friendly and open approach to start the conversation. [...] Your questions are open-ended, which encourages Jessica to talk about her situation. You have also acknowledged her feelings and concerns, creating a trusting atmosphere. [...] One area where you could possibly improve is to show empathetic concern earlier in the conversation when it comes to the topic of drinking behavior and alcohol consumption, and gently point out potential risks to Jessica. This could raise her awareness of the problem without confronting her too much. Furthermore, when it comes to creating a study plan or finding solutions for her stress situation, you could work with Jessica to develop concrete steps together, instead of giving suggestions. This would strengthen her autonomy and sense of responsibility. In conclusion, I find it very good that you offer a follow-up appointment, thus ensuring continuity in support. This gives Jessica the feeling of not being alone and being supported. Overall, you have shown a supportive and encouraging counseling attitude that is certainly helpful for Jessica. Keep it up!

Examining the differences in model performance with specific feedback methods provides insights into how structured approaches can influence feedback quality across both emotional and objective evaluations. This analysis is crucial for understanding which methods enhance or hinder the feedback generated by different models. These findings are visually represented in the heatmaps in Figure 3, with the middle heatmap showing emotional evaluation differences and the right heatmap showing objective evaluation differences.

The GPT-4 model is consistent across various methods but underperforms with the STATE method in both emotional and objective evaluations (0.5 below GPT-4 emotional rating av-

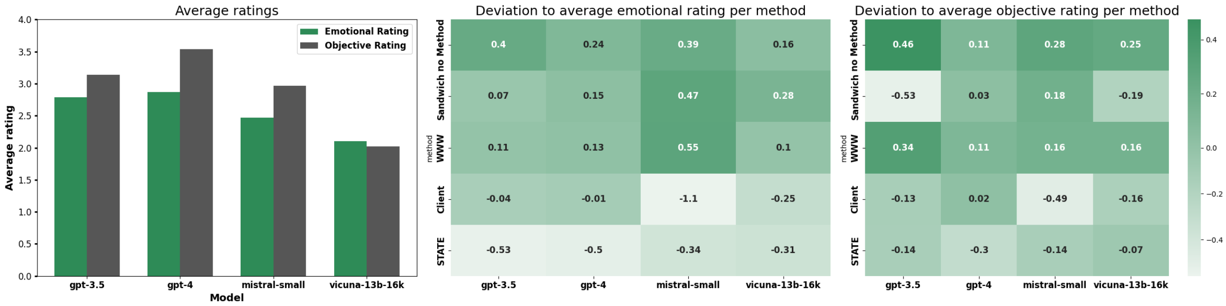

Fig. 3. Average ratings per model (left) and model performance per method (middle and right). Each column in the heatmaps contains the deviations per method to the average rating shown left for a specific model. Positive values indicate that a method works relatively well for the model.

erage of 2.87 and 0.3 below objective rating average of 3.54). In both emotional and objective evaluations, GPT-3.5 and Mistral-small perform best without a specific method (between 0.28 and 0.46 above respective model's average), indicating their strength in generating feedback without constraints. Vicuna-13b consistently performs poorly, with the client and STATE method causing the largest negative deviation from its average score.

Specifically, the GPT-3.5 model significantly underperforms with the Sandwich method in the objective evaluation, showing a decrease of -0.53. The Mistral-small model experiences the highest overall deterioration with the client method, showing a decline of -1.1 in the emotional evaluation and -0.49 in the objective evaluation. Structured methods like Sandwich and WWW generally enhance performance across most models, reflecting their utility in promoting clear and constructive feedback. However, the STATE method often results in lower scores, suggesting it may be less effective.

Overall, while certain feedback methods can improve feedback quality, as shown in the heatmaps in Figure 3, the effectiveness of each method varies significantly across models. Understanding these nuances is essential for optimizing feedback methodologies to suit each model's strengths.

B. Evaluation of Feedback Length

In Figure 4, the distribution of rated feedback lengths is illustrated. For the "Too long" category, the distribution is more spread out, with a higher concentration around mid-lengths (approximately 200-400 words). The "slightly too long" category, where the first and second epochs show disagreements between "Too long" and "Perfect Length" displays a higher density around 100-200 words but is more dispersed compared to other categories.

The "perfect length" length category has a more pronounced peak around 100-190 words, indicating a higher concentration of feedback perceived as the perfect length within this range. The "slightly too short" category exhibits a smaller range, with most feedback lengths clustering around 100 words or less. As expected, there is some overlap between these classes, reflecting the variability in feedback length depending on various factors. These factors include the feedback giver and

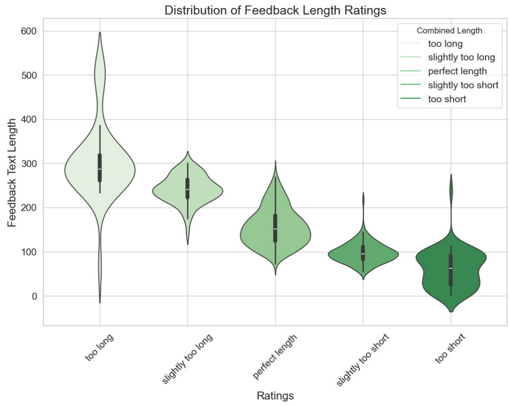

Fig. 4. Distribution of feedback text length ratings

receiver, the complexity and length of the counseling, as well as the quality of the counseling (e.g., criticism).

The distribution of length ratings, as depicted in Table I, offers insights into the perceived adequacy of generated content across different models. Notably, Vicuna-13b garnered the highest count of "too short" and "slightly too short" ratings, suggesting a tendency towards generating shorter responses compared to other models. Conversely, gpt-4 received the most "too long" and "slightly too long" ratings, indicating a propensity for verbosity in its outputs. Interestingly, despite variations in performance across other metrics, both gpt-3.5 and mistral-small received only one "too long" rating each, hinting at potential consistency in their generation of content length. The predominance of "perfect length" ratings across all models suggests a general satisfaction with the length of generated feedbacks.

TABLE I
FREQUENCY OF LENGTH RATINGS FOR DIFFERENT MODELS

length rating	**GPT-3.5**	**GPT-4**	**Mistral**	**Vicuna**	**total**
too short	6	1	18	39	**64**
slightly too short	51	1	31	45	**128**
exactly right	150	144	150	103	**547**
slightly too long	2	41	10	11	**64**
too long	1	23	1	12	**37**

When examining the correlation between feedback length rating and emotional and objective ratings, several patterns emerge, as shown in figure 5. Feedback perceived as “too short” consistently receives lower emotional (1.58) and objective (1.34) ratings. In contrast, feedback rated as “perfect length” scores the highest for both emotional (2.71) and objective (3.16) aspects. This suggests that there is an optimal range for feedback length, where it is considered thorough yet concise enough to be effective.

Interestingly, while "too long" feedback tends to score lower than "perfect length" feedback, it still fares better than "too short" feedback. This indicates that verbosity is less detrimental to perceived feedback quality than brevity. Thus, while concise feedback is ideal, slightly longer feedback is preferable to overly brief comments. Additionally, the higher ratings for longer feedback can be attributed to GPT-4, which generally provides longer feedback but consistently receives good ratings, as previously established.

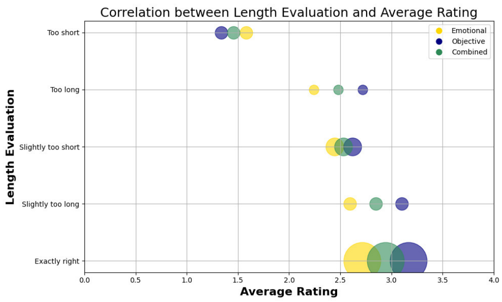

Fig. 5. Correlation between Length Evaluation and Average Rating

C. Fine-Tuning

Figure 7 in the appendix illustrates the evaluation and training loss throughout the fine-tuning process of the Vicuna-13b-16k model. The model demonstrates improvement on the evaluation dataset up to step 88, after which the evaluation loss begins to rise. This increase may suggest overfitting to the training data, as the training loss continues to decrease until the end. Consequently, we used a checkpoint at step 122, since no checkpoint was made at step 88.

The bar chart in figure 6 compares the emotional and objective scores of the Vicuna model and its fine-tuned version across different categories: ’No Method’, ’Client’, ’Sandwich’, ’STATE’, and ’WWW’. It also includes an ’Average’ category to summarize overall performance. For each category pair, the emotional scores are represented by the bars on the left, while the objective scores are represented by the bars on the right. The changes due to fine-tuning are overlaid with either green (indicating improvement) or red (indicating deterioration) bars.

The Vicuna model received an average emotional rating of 2.37 and an objective rating of 2.6. In contrast, the fine-tuned model (vicuna-ft) achieved higher ratings in both dimensions, with an emotional rating of 2.6 and an objective rating of around 3.29. This indicates that the fine-tuned model outperforms the original Vicuna model in both emotionality

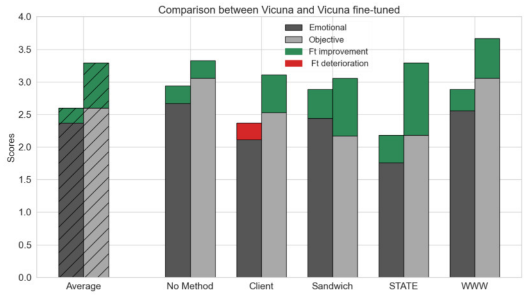

Fig. 6. Comparison of Emotional and Objective ratings between Vicuna and Fine-Tuned Vicuna across different feedback methods

(improvement of 0.23) and objectivity (improvement of 0.69). The gap between the improvements of emotionality and objectivity could be through the training data, since the training data had also higher ratings for objectivity. It shows that the fine-tuned version consistently outperforms the Vicuna model in both emotional and objective ratings, with the exception of emotionality for the category “Client”.

According to sample checks, the feedback from the fine-tuned Vicuna variant seems to contain more continuous text and focuses more often on the essentials of the counseling. Nevertheless, mistakes are often made and feedback from GPT-4, for example, still seems to to provide feedback that is much more conducive to learning. An example of feedback from the fine-tuned Vicuna model is shown below:

Sample Feedback from fine-tuned Vicuna

I noticed that you created an open and warm atmosphere at the beginning of the conversation by giving Luisa the opportunity to speak to you in the way she wanted. You also quickly made a connection with Luisa by asking about her experience and making her feel that you understood her. My impression was that you listened well and encouraged Luisa to share her feelings. You also tried to encourage Luisa to have the conversation with her parents and helped her to develop a strategy to do this. I would appreciate it if in the future you could be even more responsive to Luisa’s needs and feelings and help her to develop concrete steps to improve her situation. It would also be helpful if you could encourage Luisa to recognize and use her own strengths and resources to cope with her situation. Overall, you did a good job of supporting Luisa and making her feel understood and taken seriously.

D. Comparison of Peer-Feedback and AI based Feedback

The analysis of the sample feedbacks shows that all feedback used positive and appreciative language. A noticeable difference is the length of the feedback: peer feedback was significantly shorter and often more general and vague. For instance, general statements like *"…I found your chat very* *helpful…"* were common, with little specific observation. This could be due to the short processing time of the feedback providers and the nature of the test task.

In some cases, peer feedback included specific comments based on the conversation context, but these were often only related to the initial phase of the chat counseling. These comments were positive but not very detailed: *"…your counseling was well done. I especially liked your introduction with the framework conditions."*

In contrast, the AI-generated feedback was more detailed and specific. It highlighted the importance of the framework conditions and provided constructive criticism and suggestions for improvement: *"Additionally, you asked Lars what state he wants to achieve in 3-4 weeks, which shows forward-thinking. It would have been helpful if you had also helped him set smaller, achievable goals to support him on his journey and give him a sense of progress."*

The AI feedback also included concrete examples that could be applied in future exercises: *"For example, you could say: 'It sounds like you're going through a really challenging time. Would you like to tell me more about how you feel?' "*

The AI feedback was structured and presented clearly according to the given prompts, which can be a double-edged sword. Highly motivated students might get demotivated by constantly receiving new improvement suggestions. Therefore, feedback should come from various sources, including human feedback from course instructors and other students. This provides a mix of professional feedback and valuable perspectives from peers.

The quality of peer feedback could be improved through targeted exercises for the students. Additionally, students should be able to critically reflect on the AI feedback to avoid biases in their perceptions. A significant advantage of AI-generated feedback is that it can be requested promptly after an exercise, providing immediate feedback and eliminating the waiting time.

VII. Limitations and Future Work

While this study provides valuable insights into the implementation and evaluation of AI-generated feedback in counselor education, several limitations should be acknowledged. These limitations also suggest directions for future research and practice.

It is important to acknowledge that GPT-3.5 and GPT-4 are closed-source language models, raising ethical concerns about their use in automatic feedback generation. Furthermore, XAI approaches cannot be applied to closed-source LLMs, and there is no guarantee regarding the fate of the data involved.

The study's findings are based on a sample of 64 conversations in the context of addiction and family counseling, which, although valuable, may not fully represent the wide variety of scenarios encountered in counselor education. Future research with larger and more diverse datasets could help validate these findings and enhance their applicability across different educational contexts and counseling situations. The analysis comparing AI-generated feedback to peer feedback involved only 13 instances, which may not capture the full spectrum of feedback quality and styles. Future research with larger and more diverse datasets could help validate these findings and enhance their applicability across different educational contexts and counseling situations.

Our evaluation focused primarily on emotional and objective criteria, as well as feedback length. While these metrics offer important insights on feedback quality, they do not capture all dimensions of effective feedback, such as specificity of improvement suggestions. Expanding the range of evaluation metrics in future studies could provide a more comprehensive understanding of AI-generated feedback's impact.

The immediate reception and perceived quality of AI-generated feedback were the primary focus of this study. Future research should explore the long-term effects of such feedback on learners' skill development and overall learning outcomes. Longitudinal studies tracking progress over time would offer valuable insights into the sustained benefits and potential limitations of integrating AI feedback into counselor education.

This study utilized specific AI models, including GPT-3.5, GPT-4, Vicuna-13b, and Mixtral-8x7b1. While these models represent a range of capabilities, the findings may not be directly applicable to other AI models or future iterations. Broadening the scope to include a wider array of models in future research will help ensure the findings' relevance as AI technology evolves.

Future work should also delve deeper into the development of feedback-mechanisms for specific counseling methods and feedback at the utterance level. This could enhance the granularity and relevance of AI-generated feedback. Additionally, research should focus on longitudinal studies to assess the long-term impact of AI feedback on counselor education. Exploring ways to integrate AI feedback seamlessly with human feedback will also be crucial to develop a balanced and effective training ecosystem. Furthermore, addressing ethical considerations in greater depth, will be essential as AI becomes more integrated into educational settings.

By acknowledging these limitations and outlining directions for future research, we aim to provide a pathway for continued improvement in the effectiveness and ethical implementation of AI-generated feedback in counselor education.

VIII. Conclusion

This study explored the implementation and evaluation of AI-generated feedback in an ITS designed for counselor education. The findings demonstrate that AI models can generate detailed and specific feedback that meets objective criteria effectively. However, challenges remain in ensuring that the feedback resonates emotionally with learners, especially when using smaller open source models. The research questions described in section I are answered concisely below.

AI models can indeed provide suitable feedback on both an emotional and objective level in text-based counseling sessions, though emotional resonance remains an area for improvement. The optimal length for LLM-generated feedback tends to be between 100-190 words. Feedback that is perceived as “too short” often receives lower emotional and objective ratings, whereas feedback considered “too long” is less detrimental but still less effective than feedback of optimal length.

Different feedback methods such as Sandwich, WWW, and STATE do not consistently improve the quality of AI-generated feedback. Some models perform better without the constraints of a specific feedback method, indicating that while structured approaches can offer clarity, they may not always enhance effectiveness. The study also found that AI-generated feedback tends to score higher in objective evaluation compared to emotional evaluation, suggesting a need for improvements in emotional resonance.

Fine-tuning can enhance the performance of LLMs, as evidenced by improvements in both emotional and objective ratings for the Vicuna model after fine-tuning. This indicates that targeted fine-tuning can effectively enhance the empathetic and objective aspects of feedback generation. AI-based feedback is generally accepted, with higher acceptance among individuals who have prior experience using AI systems. Surveys indicate that familiarity with AI technologies positively influences attitudes toward AI-generated feedback, suggesting that exposure and education can enhance acceptance.

AI-generated feedback was found to be more detailed and actionable compared to peer feedback, highlighting its potential to supplement traditional feedback mechanisms in counselor education. Despite the promising results, the study’s limitations include a relatively small and specific sample size, a narrow scope of evaluation metrics, a small number of survey participants, and the limited analysis of only 13 instances of peer feedback. These limitations suggest several directions for future research, including the need for larger and more diverse datasets, a broader range of evaluation metrics, deeper exploration of ethical issues, and longitudinal studies to assess the long-term impacts of AI feedback.

Potential risks associated with AI-based feedback include the propagation of social biases present in training data and the possibility of AI models “hallucinating” incorrect or misleading feedback. Mitigation strategies include integrating multiple feedback options, continuous evaluation and fine-tuning of models, and ensuring transparency and accountability in AI-based feedback systems. Additionally, the use of a mentored control instance, where feedback generated by AI is reviewed and moderated by a human expert, can help mitigate these risks by providing an additional layer of oversight and ensuring that the feedback remains accurate and contextually appropriate.

In conclusion, while AI-generated feedback offers a powerful tool for counselor education, its implementation must be approached with careful consideration of its limitations and ethical implications. By addressing these areas in future research, we can ensure that AI systems effectively support the development of counseling competencies and contribute positively to the educational experience.

TABLE II
PARAMETER SETTINGS FOR FINE-TUNING

Parameter	**Value**
Sequence Length	4096
Sample Packing	True
Pad to Sequence Length	True
Eval Sample Packing	False
Adapter Type	LoRA
LoRA Rank (r)	32
LoRA Alpha	16
LoRA Dropout	0.05
LoRA Target Linear	True
LoRA Fan In/Fan Out	Enabled
Gradient Accumulation Steps	4
Micro Batch Size	2
Number of Epochs	6
Optimizer	AdamW
Learning Rate Scheduler	Cosine
Learning Rate	0.0002
Flash Attention	True
Eval Max New Tokens	256

APPENDIX

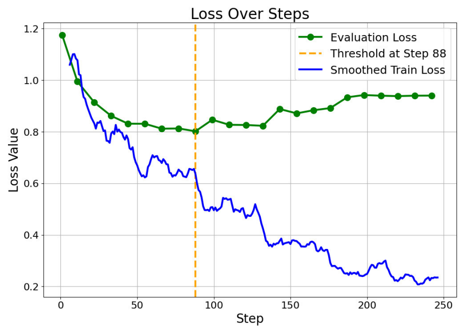

Fig. 7. Training and Evaluation Loss during the fine-tuning process with 6 epochs

REFERENCES

[1] J. Hattie and H. Timperley, “The Power of Feedback,” *Review of Educational Research*, vol. 77, no. 1, pp. 81–112, Mar. 2007. doi: 10.3102/003465430298487

[2] S. Narciss, S. Sosnovsky, L. Schnaubert, E. Andrès, A. Eichelmann, G. Goguadze *et al.*, “Exploring feedback and student characteristics relevant for personalizing feedback strategies,” *Computers & Education*, vol. 71, pp. 56–76, Feb. 2014. doi: 10.1016/j.compedu.2013.09.011

[3] W. Dai, J. Lin, H. Jin, T. Li, Y.-S. Tsai, D. Gašević *et al.*, “Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT,” in *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, Jul. 2023. doi: 10.1109/ICALT58122.2023.00100 pp. 323–325, iSSN: 2161-377X.

[4] R. E. Wang, Q. Zhang, C. Robinson, S. Loeb, and D. Demszky, “Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes,” Apr. 2024, arXiv:2310.10648 [cs].

[5] E. M. Engelhardt, “Onlineberatung – Digitales Beratungsangebot für Alle?” in *Digital Diversity*, H. Angenent, B. Heidkamp, and D. Kergel, Eds. Wiesbaden: Springer Fachmedien Wiesbaden, 2019, pp. 161–173. ISBN 978-3-658-26752-0 978-3-658-26753-7

[6] M. Stieler, S. Lipot, and R. Lehmann, “Zum Stand der Onlineberatung in Zeiten der Corona Krise. Entwicklungs- und Veränderungsprozesse der Onlineberatungslandschaft,” *e-beratungsjournal.net – Zeitschrift für Online-Beratung und computervermittelte Kommunikation*, vol. 18,

no. 1, pp. 50–65, 2022. doi: 10.48341/262P-7T64 Publisher: Universität für Weiterbildung Krems & e-beratungsjournal.net.

[7] A. Chaszczewicz, R. S. Shah, R. Louie, B. A. Arnow, R. Kraut, and D. Yang, "Multi-Level Feedback Generation with Large Language Models for Empowering Novice Peer Counselors," Mar. 2024, arXiv:2403.15482 [cs].

[8] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain *et al.*, "Large Language Models: A Survey," Feb. 2024, arXiv:2402.06196 [cs].

[9] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang *et al.*, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," Nov. 2023, arXiv:2311.05232 [cs].

[10] E. Mousavinasab, N. Zarifsanaiey, S. R. Niakan Kalhori, M. Rakhshan, L. Keikha, and M. Ghazi Saeedi, "Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods," *Interactive Learning Environments*, vol. 29, no. 1, pp. 142–163, Jan. 2021. doi: 10.1080/10494820.2018.1558257

[11] E. Rudolph, N. Engert, and J. Albrecht, "An AI-Based Virtual Client for Educational Role-Playing in the Training of Online Counselors," in *Proceedings of the 16th International Conference on Computer Supported Education - Volume 2: CSEDU*, vol. 2. SCITEPRESS, May 2024. doi: 10.5220/0012690700003693. ISBN 978-989-758-697-2 pp. 108–117.

[12] J. Carbonell, "AI in CAI: An Artificial-Intelligence Approach to Computer-Assisted Instruction," *IEEE Transactions on Man-Machine Systems*, vol. 4, no. 11, pp. 190–202, 1970. doi: 10.1109/TMMS.1970.299942

[13] A. T. Corbett, K. R. Koedinger, and J. R. Anderson, "Chapter 37 - Intelligent Tutoring Systems," in *Handbook of Human-Computer Interaction (Second Edition)*, M. G. Helander, T. K. Landauer, and P. V. Prabhu, Eds. Amsterdam: North-Holland, Jan. 1997, pp. 849–874. ISBN 978-0-444-81862-1

[14] W. Ma, O. O. Adesope, J. C. Nesbit, and Q. Liu, "Intelligent tutoring systems and learning outcomes: A meta-analysis," *Journal of Educational Psychology*, vol. 106, no. 4, pp. 901–918, 2014. doi: 10.1037/a0037123 Place: US Publisher: American Psychological Association.

[15] C. Cao, "Scaffolding CS1 Courses with a Large Language Model-Powered Intelligent Tutoring System," in *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, ser. IUI '23 Companion. New York, NY, USA: Association for Computing Machinery, 2023. doi: 10.1145/3581754.3584111. ISBN 9798400701078 pp. 229–232.

[16] J.-Y. Kuo, H.-C. Lin, P.-F. Wang, and Z.-G. Nie, "A Feedback System Supporting Students Approaching a High-Level Programming Course," *Applied Sciences*, vol. 12, no. 14, p. 7064, Jan. 2022. doi: 10.3390/app12147064 Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

[17] Z. Marafie, K.-J. Lin, D. Wang, H. Lyu, Y. Liu, Y. Meng *et al.*, "AutoCoach: An Intelligent Driver Behavior Feedback Agent with Personality-Based Driver Models," *Electronics*, vol. 10, no. 11, p. 1361, Jan. 2021. doi: 10.3390/electronics10111361 Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[18] A. Botelho, S. Baral, J. A. Erickson, P. Benachamardi, and N. T. Heffernan, "Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics," *Journal of Computer Assisted Learning*, vol. 39, no. 3, pp. 823–840, 2023. doi: 10.1111/jcal.12793

[19] B. Grawemeyer, M. Mavrikis, W. Holmes, S. Gutierrez-Santos, M. Wiedmann, and N. Rummel, "Affecting off-task behaviour: how affect-aware feedback can improve student learning," in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, ser. LAK '16. New York, NY, USA: Association for Computing Machinery, Apr. 2016. doi: 10.1145/2883851.2883936. ISBN 978-1-4503-4190-5 pp. 104–113.

[20] M. Dzikovska, N. Steinhauser, E. Farrow, J. Moore, and G. Campbell, "BEETLE II: Deep Natural Language Understanding and Automatic Feedback Generation for Intelligent Tutoring in Basic Electricity and Electronics," *International Journal of Artificial Intelligence in Education*, vol. 24, no. 3, pp. 284–332, Sep. 2014. doi: 10.1007/s40593-014-0017-9

[21] J. McDonald, A. Knott, S. Stein, and R. Zeng, "An empirically-based, tutorial dialogue system: design, implementation and evaluation in a first year health sciences course." in *Proceedings of Electric Dreams. Proceedings ascilite 2013*. Australasian Society for Computers in Learning in Tertiary Education, 2013. ISBN 978-1-74138-403-1 pp. 562–572.

[22] F. St-Hilaire, D. D. Vu, A. Frau, N. Burns, F. Faraji, J. Potochny *et al.*, "A New Era: Intelligent Tutoring Systems Will Transform Online Learning for Millions," Mar. 2022.

[23] C. Cao, "Leveraging Large Language Model and Story-Based Gamification in Intelligent Tutoring System to Scaffold Introductory Programming Courses: A Design-Based Research Study," Feb. 2023, arXiv:2302.12834 [cs].

[24] Y. Y. Chiu, A. Sharma, I. W. Lin, and T. Althoff, "A Computational Framework for Behavioral Assessment of LLM Therapists," Jan. 2024, arXiv:2401.00820 [cs].

[25] M. Glickman and T. Sharot, "How human-AI feedback loops alter human perceptual, emotional and social judgements," Nov. 2022.

[26] J. Lin, L. Sha, Y. Li, D. Gasevic, and G. Chen, "Establishing Trustworthy Artificial Intelligence in Automated Feedback," Jul. 2022.

[27] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy Artificial Intelligence: A Review," *ACM Computing Surveys*, vol. 55, no. 2, pp. 39:1–39:38, Jan. 2022. doi: 10.1145/3491209

[28] D. Gursoy, O. H. Chi, L. Lu, and R. Nunkoo, "Consumers acceptance of artificially intelligent (AI) device use in service delivery," *International Journal of Information Management*, vol. 49, pp. 157–169, Dec. 2019. doi: 10.1016/j.ijinfomgt.2019.03.008

[29] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003. doi: 10.2307/30036540 Publisher: Management Information Systems Research Center, University of Minnesota.

[30] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," Jul. 2023, arXiv:2307.09288 [cs].

[31] H. Luo and L. Specia, "From Understanding to Utilization: A Survey on Explainability for Large Language Models," Feb. 2024, arXiv:2401.12874 [cs].

[32] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt *et al.*, "Bias and Fairness in Large Language Models: A Survey," Mar. 2024, arXiv:2309.00770 [cs].

[33] N. Dainton, "1 Bedeutung und Wert von Feedback," in *Feedback in der Hochschullehre*, ser. utb-Titel ohne Reihe. Haupt, Dec. 2020, pp. 11–22. ISBN 978-3-8252-4891-8

[34] J. Fengler, *Feedback geben: Strategien und Übungen; ... mit über 100 Übungen*, 4th ed., ser. Beltz Weiterbildung. Weinheim Basel: Beltz, 2009. ISBN 978-3-407-36471-5

[35] R. de Villiers, "7 Principles of highly effective managerial feedback: Theory and practice in managerial development interventions," *The International Journal of Management Education*, vol. 11, no. 2, pp. 66–74, Jul. 2013. doi: 10.1016/j.ijme.2013.01.002

[36] J. Ade and U. Gläßer, "Lehrmodul 12: Feedback in der Mediation," *Zeitschrift für Konfliktmanagement*, vol. 12, no. 2, Jan. 2009. doi: 10.9785/ovs-zkm-2009-60

[37] N. Dainton, "2 Wo klemmt es," in *Feedback in der Hochschullehre*, ser. utb-Titel ohne Reihe. Haupt, Dec. 2020, pp. 23–32. ISBN 978-3-8252-4891-8

[38] A. Seidl, "Dein Wunsch geht in Erfüllung," in *Freundlich, aber bestimmt – Die richtigen Worte finden in Gesundheitsberufen*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 51–70. ISBN 978-3-642-41803-7 978-3-642-41804-4

[39] A. Dohrenwend, "Serving Up the Feedback Sandwich," *Family Practice Management*, vol. 9, no. 10, pp. 43–46, Nov. 2002.

[40] A. J. Henley and F. D. DiGennaro Reed, "Should You Order the Feedback Sandwich? Efficacy of Feedback Sequence and Timing," *Journal of Organizational Behavior Management*, vol. 35, no. 3-4, pp. 321–335, Oct. 2015. doi: 10.1080/01608061.2015.1093057

[41] J. Sammet and J. Wolf, "Präsenztraining im Blended Learning," in *Vom Trainer zum agilen Lernbegleiter: So funktioniert Lehren und Lernen in digitalen Zeiten*, J. Sammet and J. Wolf, Eds. Berlin, Heidelberg: Springer, 2019, pp. 55–65. ISBN 978-3-662-58510-8

[42] J. Grenny, K. Patterson, R. McMillan, A. Switzler, and E. Gregory, *Crucial Conversations*, 3rd ed. New York: McGraw Hill, 2022. ISBN 978-1-260-47419-0

[43] F. Eidenbenz, "Standards in der Online-Beratung," in *Handbuch Online-Beratung*. Vandenhoeck & Ruprecht GmbH & Co. KG, Nov. 2009, pp. 213–228. ISBN 978-3-525-40154-5