Can GPT-4 Enhance Teaching? A Pilot Study on AI-Driven Analysis of Student Course Feedback

Oshani Weerakoon[1], Panu Puhtila[1], Tuomas Mäkilä[1] and Erkki Kaila[1]

[1]*University of Turku, Turku, Finland*

Abstract

In this pilot study, we explored the use of generative AI—specifically GPT-4—to evaluate student feedback in a bilingual software engineering course offered at the University of Turku. Our aim was twofold: to examine whether ChatGPT can meaningfully evaluate student course feedback and propose suitable enhancements, and to compare its evaluations with those made by a course teacher. We collected voluntary feedback from 18 consenting students across three course instances in 2023 and 2024, resulting in a total of 390 feedback entries. These responses were first translated into English and then anonymized. Using structured questionnaires aligned with defined pedagogical goals, we then analyzed the responses through a dual evaluation process: (1) AI-based assessment using a custom JavaScript application integrating GPT-4 and GPT-4o-mini, and (2) manual evaluation by the teacher. Both followed a standardized Likert-scale format with brief textual comments, and all evaluations were consolidated into thirty-six manually maintained recording sheets. Evaluation results were visualized using heat maps across five key themes derived from the pedagogical goals. Our comparative analysis showed general alignment between the two evaluators, with key differences in the perceived content clarity and video quality of the course. We further extended our discussion to examine GPT's applicability and limitations as a feedback evaluator. In particular, we identified its potential to quickly assess structured student feedback in courses with high participation, where manual evaluation may be time-consuming for course teachers. These findings collectively provide insights into using generative AI in course feedback analysis to enhance teaching within software engineering curricula.

Keywords

Generative AI, Engineering Education, Pedagogy, GPT

1. Introduction

The swift development of artificial intelligence technologies in recent years has significantly improved their usefulness across several industries. Notably, the educational industry has received a lot of attention, especially given the potential uses of artificial intelligence in teaching and learning.

ChatGPT is a Large Language Model that was released in November 2022 by OpenAI [1]. It is based on the Generative Pretrained Transformer (GPT) architecture [2], which uses deep learning techniques to generate human-like text. ChatGPT models, including GPT-3.5, and GPT-4, are trained on vast amounts of text data and leverage transformer architectures, featuring self-attention mechanisms that allow them to understand context, generate coherent responses, and complete tasks like translation, summarization, and conversation. GPT-4o is OpenAI's latest flagship model, capable of real-time reasoning spanning voice, vision, and text [3]. The current context window for GPT-4o is 128k, with a knowledge cut-off date of October 2023 [3]. These models excel at tasks requiring language understanding and generation, making them useful in chatbots, content creation, and coding assistance [2].

Evaluations of teaching and learning by students themselves have become progressively vital in facilitating high-quality, learner-centric education [4]. Gaining insight into how students see their educational experiences provides important information that may affect course design and improvement. Many feedback tools can be integrated into the course structure to collect feedback from the students. However, the results may generally be summarised in the form of charts or it will be time-consuming for teachers to go through and derive detailed insights when the number of student feedback is huge.

TKTP 2025: Annual Doctoral Symposium of Computer Science, 2.—3.6.2025 Helsinki, Finland

osweer@utu.fi (O. Weerakoon); papuht@utu.fi (P. Puhtila); tusuma@utu.fi (T. Mäkilä); ertaka@utu.fi (E. Kaila)

0009-0004-1684-239X (O. Weerakoon); 0009-0004-6418-1063 (P. Puhtila); 0000-0002-8799-185X (T. Mäkilä); 0000-0002-2407-9492 (E. Kaila)

ChatGPT has shown capability in simplifying time-consuming text-based analysis and summarizing tasks owing to its massive set of trained data and transformer architecture. It further demonstrates various intelligent behaviours which might even closely follow human writing [5]. Therefore, this study aims to explore the usage of GPT-4 as a novel workflow to derive meaningful evaluations of course feedback given by students and comparing with those made by the teachers.

While there are several Generative AI tools available in the market, we chose GPT-4 for this study for two reasons. Firstly, ChatGPT is easily accessible and is widely used by many people including teachers who are our target audience from this research. In this research, we offer an API-based web interface to evaluate student feedback, however, ChatGPT's conversational interface can easily be used to perform the evaluation task.

On the other hand, OpenAI's GPT models are still leading a lot of Generative AI-based research in different fields and OpenAI has been continuing to upgrade the models to fit specific tasks like reasoning, coding etc. which makes the use cases of the model still high among its competitors.

While Generative AI tools like GPT-4 can be extremely useful in various applications, there are instances where they produce text that may be perceived by humans as incorrect or deceptive. These instances, known as hallucinations [6], [7], [8] are situations where the responses seem logical or consistent, despite being factually incorrect. Also, there have been recent findings about ChatGPT using "shortcuts" to generate text, rather than deep reasoning [9]. The privacy and security issues related to working with ChatGPT are also a concern that could not be overlooked [10]. Furthermore, the information generated in ChatGPT can be incorrect and biased [11], which emphasizes the necessity of handling the AI-generated content after careful review.

In light of these developments and known limitations, our study aims to answer the following research questions:

1. **RQ1**: Can ChatGPT evaluate student course feedback and propose suitable enhancements?
2. **RQ2**: How do ChatGPT's evaluations compare to those of a teacher?

We present in this paper a novel workflow utilizing GPT-4 API and GPT-4o-mini API for evaluating feedback given by students for a software engineering course held at the University of Turku. We then compare the evaluations made by GPT against those made by a course teacher to assess the efficacy of the AI. We employed both qualitative and quantitative analyses to explore answers to the research questions posed above.

Our paper is outlined as follows. Section 2 explains related studies and Section 3 explains the methodology that we used to conduct this research in detail. Section 4 presents and compares the evaluations made by ChatGPT and the teacher on the student feedback. Section 5 extends the discussion and answers the research questions. Section 6 concludes the findings of our study and proposes future work.

2. Related Work

Since the body of research focusing on the specific phenomena we are studying is quite sparse, we will here discuss studies that are contextually and thematically adjacent to our work. These include various research that has investigated the factors influencing the quality of feedback, as well as a few studies done on the usage of Generative AI tools in giving feedback to students. We will present them in this order.

QUEST was an assessment tool designed by Tricker et al. [12] to gauge student satisfaction with distance education courses within the higher education system. It has produced metrics assessing different aspects of the course's service quality which can enhance the alignment between student expectations and their actual experiences in the course.

Brew [13] investigated in her study whether or not students would respond in-depth to an anonymous survey, whether or not the survey's integration into the course management system was successful, and whether or not the responses would be helpful criticism for course assessment and changes. 71

people responded to the survey, which was administered in 2001, 2003, 2005, and 2007. The survey results showed students gave thoughtful and comprehensive comments with more success when the student count was low.

Dai et al. [14] explored the potential of utilizing ChatGPT as a tool for delivering feedback to enhance student learning. The findings indicated that: (i) ChatGPT was able to produce feedback that was not only more detailed but also fluently captured students' performance compared to human instructors; (ii) there was a high level of agreement between ChatGPT and the instructors in evaluating the subject matter of students' assignments; and (iii) ChatGPT was capable of providing insights into the methods students employed to complete their tasks, potentially aiding in the development of their learning skills. However, their research did not investigate utilizing generative AI tools to evaluate student feedback to improve course teaching.

Another similar study, conducted by Dai et al. [15] examines the potential of GPT models in enhancing AI-driven learning analytics for assessment feedback. It evaluates GPT-3.5 and GPT-4 in generating feedback for student writing tasks in a data science course, comparing them with human instructors in readability, effectiveness, and reliability. Results show that GPT models, particularly GPT-4, produce more readable and consistent feedback, outperform human instructors in providing structured feedback, and demonstrate higher reliability.

A study done by Morales-Chan et al. [16] also explores AI-driven feedback automation in MOOCs to enhance student learning. By combining LangChain [1] and the OpenAI API, the system generates personalized, rubric-based feedback aligned with educational goals. More than just automating responses, this approach reshapes digital education by making MOOCs more interactive and adaptive. The research also highlights generative AI's potential to scale individualized feedback, improving learning experiences globally. Findings indicate improved student satisfaction and progress, confirming AI's value in self-paced courses.

Utilizing Llama2 (7B) to summarize 742 student course feedback responses across 75 Computer Science courses, Zhang et al. [17] conducted a study to assess the capability of Llama2 to provide course instructors with actionable insights. Findings suggest that generative AI can effectively synthesize factual and relevant feedback, offering a cost-efficient solution for enhancing teaching practices. However, no comparison study with human evaluators has been made in this study. Despite this, the study demonstrates the feasibility of AI-driven evaluation summaries to support educators' professional development, highlighting a promising approach to improving teaching feedback in large classroom settings.

In the study conducted by Steiss et al. [18], they investigated the capacity of generative AI (ChatGPT) to offer formative feedback and compared the quality of feedback provided by humans and AI. Their analysis included 200 instances of formative feedback generated by humans and 200 by AI for identical essays. The evaluation showed that feedback from well-trained evaluators was of higher quality compared to that from ChatGPT. However, given ChatGPT's ability to produce feedback easily and its reasonable quality, generative AI could still be beneficial in certain scenarios, such as when a skilled educator is not available.

Course evaluations are widely utilized in higher education, yet extracting valuable insights from numerous open-ended comments can be time-consuming. Fuller et al. [19] investigated the use of ChatGPT for analyzing course evaluation comments, focusing on the time efficiency in theme generation and the agreement between themes identified by instructors and those determined by AI. The study demonstrated ChatGPT's potential as an analytical tool for open-ended comments in health professions education. However, their results indicate that it is essential to use ChatGPT as an auxiliary tool in the analysis process and to not depend exclusively on its outputs for conclusions.

[1] https://www.langchain.com/

3. Research Methods

For this pilot study, we selected a fresh open university software engineering course, named, *Green Coding* that is conducted both in Finnish and English languages. The course is implemented under Sustainable Web Development Studies[2] category at the University of Turku, Finland.

3.1. Course Background

The course covers concepts related to designing efficient algorithms that minimize resource usage and energy consumption, as well as monitoring energy use during development and run time. Furthermore, the course encourages students to make sustainable choices, such as selecting appropriate platforms, development frameworks, and tools. The ultimate goal is fostering lower carbon footprints, balancing performance with energy efficiency, and integrating environmental concerns into software development processes.

The course comprised six chapters which took place in six different instances (separately in English and Finnish) in February 2023, June 2024 and October 2024. The course materials were available online for enrolled students to complete at their own pace, with one mandatory lecture demonstrating power consumption in software and hardware-based solutions both in Finnish and English languages. The average course duration was four weeks. These courses under the open university category are free of charge and designed for working professionals who need to update their skill sets, or in general for anyone interested in the topic.

Participants: Since the course was free and self-paced, the completion of the course depended on the students themselves. Again, providing feedback in the chapter was voluntary. Out of all the students who took the course, 18 students expressed their explicit consent to use their course feedback in this research. This includes both Finnish and English students from all three instances.

A research consent form in both English and Finnish languages was distributed among the students enrolled in the course to collect consent to process their feedback about the course. This step was imperative to ensure the ethical and consented usage of student feedback in our scientific study involving Generative AI. We further ensured the anonymity of the data by excluding any personal identifiers of 390 feedback entries from all the consented students from the analysis process.

3.2. Study Design

We collected student feedback with a form published in the Moodle instance of the course through which students were able to gauge their opinion on the course teaching, learning experience and quality of course materials. The course included six chapters, and at the end of each chapter, students could voluntarily answer the questions providing feedback about that chapter. The questionnaire was designed to capture a spectrum of pedagogical areas, as shown in Table 1. Based on these pedagogical goals, the eight feedback questions shown in Table 2 were produced. The feedback questions were also mapped to the pedagogical goals. All questions were evaluated under a 5-point based Likert scale grading criteria as follows: *1. Excellent, 2. Good, 3. Fair, 4. Poor, and 5. Needs Improvement.*

Before conducting the analysis, pre-processing of the feedback data was required. We pre-processed the 390 feedback points as below.

Data Pre-processing: This process began with exporting all the student feedback across all chapters in all three-course instances in the Excel format. Afterwards, we carried out the following steps in order: consolidating all feedback into a single dataset, excluding responses from students who did not consent to participate in the study, and removing any personal data to uphold GDPR compliance and maintain confidentiality. Given the bilingual delivery of the course in both English and Finnish, it was necessary to translate the Finnish feedback into English to optimize the performance of the ChatGPT. We utilized the DeepL [20] translation service to ensure high fidelity in our translations.

[2]https://www.utu.fi/en/open-university-studies/courses/faculty-of-technology/sustainable-web-development

Table 1
Pedagogical goals for evaluating the students' feedback

	Pedagogical Goal	For What?	Feedback Question
A	Content Understanding and Clarity	Assess new learning and unclear aspects in the course.	1, 2
B	Exercise Difficulty	Find out the difficulty ratings of the exercises to the students.	3, 4
C	Time Investment	Analyze study hours to determine pacing and course content volume issues.	5
D	Perception of AI-generated Content	Find out if students can successfully recognize AI-generated content to review the success of the course generation process with Generative AI.	6
E	Lecture Video Feedback	To know student perception about the AI-generated lecture videos to collect suggestions for improvement.	7
F	Overall Teaching Improvement	Compile suggestions for improvement across lectures and exercises.	7, 8

Table 2
The feedback questions related to the course learning and teaching

Feedback Questions
1. What did you learn new in this chapter of the course? 2. What aspects of this chapter remained unclear for you? 3. How easy did you find the exercises to attend? a. Easy b. Average c. Hard 4. Explain your rating above. 5. Estimate the number of study hours you have done to complete this chapter. a. Less than 1 hour b. 1-3 hours c. More than 3 hours 6. Do you feel any content in this chapter was generated by non-human agents like ChatGPT? If yes, explain. 7. Do you have any feedback about the lecture video of this chapter? What did you like and did not like about it? 8. How would you like to improve the teaching (lectures, exercises etc.) of this chapter?

3.3. Analysis Methods

After the student feedback data was pre-processed, we evaluated two folds:

1. GPT-4 based Evaluation
2. Teacher's Evaluation

GPT-4 based Evaluation: For the AI-based student feedback evaluation, we developed a JavaScript-based web application, 'Analyse with GPT-4' [21] that integrated the GPT-4 API (for February 2023 course instance) and GPT-4o-mini API (for June 2024 and October 2024), which is seemingly 'faster at most questions' [3] and fitted our requirement.

To provide contextual knowledge of the course and pre-defined evaluation criteria to model, the following sources of information were provided.

1. Feedback questions
2. Corresponding pedagogical goals

3. Evaluation criteria in the Likert scale
4. Summary of lecture content

All the provided content was in the form of markdown files (.md format). We analysed the feedback given for Questions 3 and 4 together to get a more subjective evaluation since these two questions are related and complement each other. Every time the model was prompted to evaluate a given student feedback, these knowledge artefacts were also sent with the prompt. Several prompts were experimented with, and the one we used was chosen because it provided the most consistent results. The prompt used is shown below.

> "Below is a feedback question and answer, given by a university student about the given chapter under the course 'X'. The chapter content and criteria for evaluating feedback per each question are given. Imagine you are the teacher of the course, refer to the chapter content first. Then, check the relevant question and its evaluation criteria. Finally, evaluate the student's answer based on the criteria and knowledge covered by the chapter. Draw useful insights to better the course but do not improvise your response beyond the given content, do not repeat the question or student's answer in the response, and generate the response with evaluation grade + less than 20 words, in plain text format. ; "

We requested each evaluation to be in the output format of *evaluation grade + further comments in plain text.* The ChatGPT response was intentionally limited to 20 words to match the length of responses the teacher had given to maintain the balance during the comparison stage of both types of evaluations.

Teacher's Evaluation: A course teacher conducted a manual evaluation of the feedback of all the students in all six chapters. The teacher's evaluation also followed the same output format as above.

The teacher and GPT-4, each evaluated a total of 390 different student feedback points across six chapters in all six course instances (both Finnish & English). Each evaluation including the grade, was separately recorded in thirty-six recording sheets manually.

Exclusion from the Analysis: In the October 2024 instance of the course, feedback Question 5- *Estimate the number of study hours you have done to complete this chapter*, was removed from the feedback form. Therefore, all the feedback received for Question 5 in other instances was also excluded from this analysis. This is mainly because the researchers identified that particular feedback cannot be evaluated objectively to a qualitative rating (Poor to Excellent) without additional information about the weight of each chapter's content.

4. Results

To get a comparative view of how the two sets of evaluations have behaved, we used heat maps to visualize the feedback evaluation grade ratings given by the teacher and AI.

4.1. Visualizing using heat maps

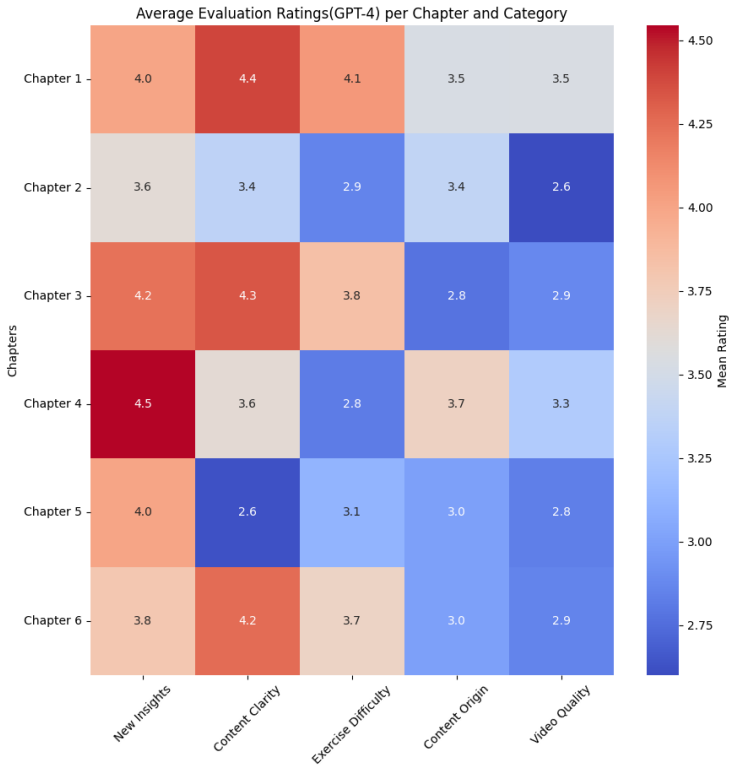
While the collected textual data (students' feedback) was qualitative, the evaluation grade being on a Likert scale allowed us to quantify these evaluations to graph appropriately. First, based on the pedagogical goals defined in Table 1, we extracted 5 themes- *New Insights, Content Clarity, Exercise Difficulty, Content Origin* and *Video Quality* that correspond to feedback questions 1, 2, 3, 4, 6 and 7 respectively. We chose the particular grades that correspond to these questions, made by the ChatGPT-4, and the teacher in the recording sheets. We then mapped them to a scale of 1-5, with 1 being *Needs Improvement* and 5 being *Excellent.*

Further, we used np.nanmean after replacing 'No answer' entries with NaN to ensure that missing or incomplete data does not unduly bias the mean estimates. Finally, we computed the mean for all numerical grades of each chapter per category, separately for the two evaluators -4 and the teacher. To visualize the evaluation data that we collected from ChatGPT-4 and the teacher, we generated two heat

maps using Seaborn and matplotlib.pyplot python libraries. The complete code used for building the heat maps is available here [22].

Figure 2 shows the heat map of the teacher's evaluation and Figure 1 shows the heat map of ChatGPT's evaluation.

Figure 1: Heatmap showing ChatGPT-4's Evaluation

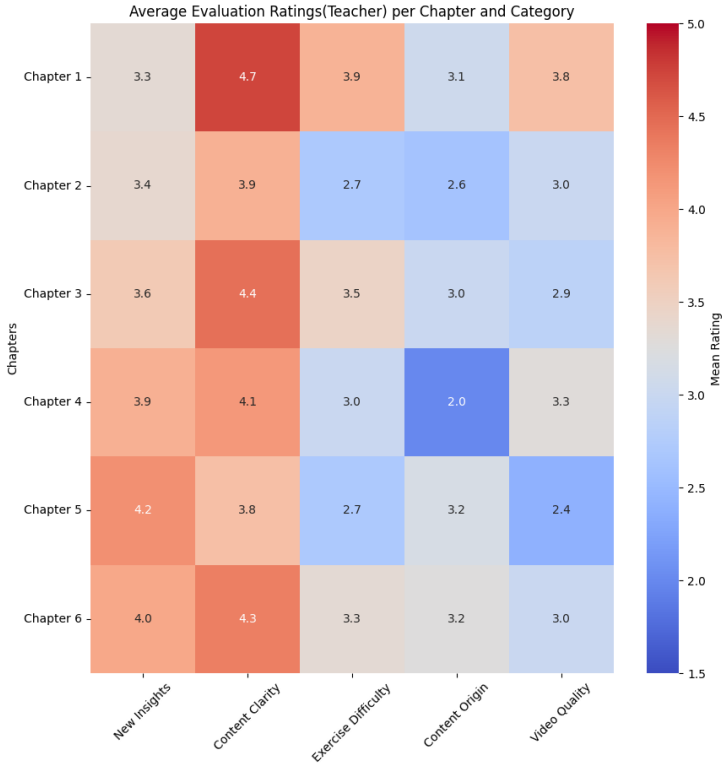


4.2. Comparison Insights

According to the heat maps generated, both ChatGPT and the teacher have given high evaluation grades for *New Insights*, indicating that they found the student responses indicative of the particular course providing the student with new knowledge. In terms of *Content Clarity*, there were some variations, but both ChatGPT and the teacher rated it from Good to Excellent, while the teacher's ratings were particularly higher than the ChatGPT's. Both the AI and the teacher recognized variability in *Exercise Difficulty*, with ratings reflecting a range of student experiences. The AI's ratings for difficulty were generally higher than the teacher's, suggesting that the teacher might perceive the exercises as less challenging compared to the AI's judgment.

In terms of *Content Origin*, the pedagogical aspect involved was unveiling student perception of AI-based lecture content generation. Both the teacher and the ChatGPT have rated the respective students' feedback as Fair to Poor. Particularly, in certain feedback answers, the students did not prefer the use of text-to-speech service in the lecture videos highlighting the "robot-like" narration. For *Video Quality*, ChatGPT provided varied ratings, and the teacher has also rated student feedback as less satisfied with the video content. This observation is also likely to stem from the preference of students towards human voice over AI-generated narration in the lecture videos.

Figure 2: Heatmap showing Teacher's Evaluation



5. Discussion

In this study, both ChatGPT (AI) and the human teacher employ an identical evaluation method characterized by a fixed set of five dimensions—New Insights, Content Clarity, Exercise Difficulty, Content Origin, and Video Quality—across six chapters of the selected course. This standardization is analogous to using a validated Likert-scale survey in quantitative research. By mapping qualitative ratings (e.g., Excellent, Good, Fair) to numerical values, our study implements subjective evaluations made by ChatGPT versus evaluations made by human teachers into objective, quantifiable data. Altogether, both evaluators examined 390 student feedback points under the structured criteria.

To answer **RQ1: Can ChatGPT evaluate student course feedback and propose suitable enhancements?**, we observed that the GPT-4 response was rather informative and complete and also suggested enhancements to course teachers. For example, *"Evaluation Grade: Good. Insights on Big O clarity highlight the importance of practical application in assignments."* is one such evaluation done by GPT on the Big O notation [3], in Chapter 5 of the course under optimization methods to increase code efficiency. It has been identified that the student's feedback reflects positively on the Big O notation in that chapter.

To answer **RQ2: How do ChatGPT's evaluations compare to those of a teacher?**, after comparing the specific colour shades and ratings in each heat map, it can be determined that the teacher's evaluations may tend to reflect a more keen understanding of student feedback, which can be due to subjective experiences and qualitative judgment. On the other hand, the AI's evaluations appear to be based on a more consistent metric-based approach, which is likely to focus on the objective criteria provided. We also found, generally, that the evaluations of ChatGPT were more informative and complete than the teacher's and tried to provide improvements to the course. We observed similarities in categories that are inherently less subjective specifically, in the *New Insights* and *Exercise Difficulty*

[3]https://web.mit.edu/16.070/www/lecture/big_o.pdf

dimensions. Both evaluators consistently rated these aspects at similarly high levels across the chapters, suggesting that the course effectively imparts new knowledge and that the exercises are appropriately challenging. Adding to this, We used two GPT-4 models for the evaluations, but we did not observe a significant change in the output in this pilot study. The subjective categories, such as *Content Clarity* and *Video Quality*, show the greatest divergence in evaluations, emphasizing the need for a more defined approach for ChatGPT to assess these aspects.

On the other hand, the related feedback on these categories was comparatively expressive, where students expressed their thoughts about AI-generated lecture videos openly-many disliking the AI-generated narration. According to Kocoń et al. [23], the accuracy of ChatGPT-4 diminishes significantly for more challenging and practical natural language processing tasks, particularly in the evaluation of texts with emotional content. Teachers can identify subtleties in student feedback, and thus assess clearly whether something has been understood, rather than just rote-learned, or if the teaching methods are engaging and effective beyond just conveying information or sentiments behind them.

Adding to this, is the potential bias in ChatGPT in qualitative analysis, likely resulting from the guidelines enforced for human trainers by OpenAI [23]. There is also the possibility that hallucinations could have affected the evaluations given by ChatGPT. Although it is impossible to completely remove the occurrence of hallucinations in ChatGPT as a language model, these can be minimized by providing ample contextual information along with the prompt [24]. Furthermore, there is the possibility of translation errors incurred by DeepL, in the Finnish feedback translated to English for having a complicated structure and/or off-meaning. It should be noted that the teacher who evaluated the feedback was not required to translate the feedback and had native proficiency in the Finnish language.

These findings confirm that while both evaluators agree on certain objective dimensions, the teacher's evaluative lens is crucial for capturing the full complexity of student feedback. The similarity in *New Insights* and *Exercise Difficulty* confirms the reliability of the measurement framework, whereas the divergence in *Content Clarity* and *Video Quality* emphasizes the need for human insight to address exact pedagogical shortcomings. It is clear that while ChatGPT can offer fast, consistent evaluations, human judgment is still required to capture the full scope of the course experience and satisfaction level of the students. Therefore, our analysis supports the dual-evaluator approach, where ChatGPT-4's consistency complements the teacher's judgment, ultimately providing a holistic overview of student feedback to improve course teaching.

Apart from the above set of 5 key themes, we continuously observed GPT as a feedback evaluator was *fast* at giving the structured evaluation with grading compared to the speed of the human teacher. When it comes to courses with a higher number of students, the speed of GPT in evaluating student feedback would greatly benefit teachers to extract insights faster and then review where needed.

Lastly, one interesting aspect that arose from the student feedback was their perception of whether the AI technologies had been used in the production of the course materials. They were used, in the production of lecture videos and slides, based on the human-written course material, and we wanted to gauge how obvious this was to the students. Our observation was that while many students pointed out, correctly, very obvious indicators for some particular presentation having been done with large Language Models(LLMs) or text-to-speech (TTS) systems, many students were completely oblivious to these elements. Even in situations where, for example, the voice in the lecture video was created with TTS technology, and every other student recognized it as being machine-generated, one student would claim that all materials seemed completely natural, that is, human-generated. This is interesting, as it points out that some people are more prone than others to not recognize AI-generated content, which is a phenomenon that would require more study in the future.

5.1. Threats of Validity

We used a standardized evaluation framework that converts qualitative student feedback into numerical data. Both ChatGPT and the human teacher applied the same Likert-scale conversion to measure the dimensions-New Insights, Content Clarity, Exercise Difficulty, Content Origin, and Video Quality. This consistent approach ensures that differences in evaluations arise from the evaluators' perspectives, not

from the measurement process, thereby ensuring the *Internal Validity*.

Before the analysis, we identified feedback questions on the number of study hours taken by students to complete a chapter, that could not be satisfactorily evaluated to a qualitative rating (e.g., Excellent, Good etc.) by the evaluators without additional information. Therefore, we did not use that particular feedback in the analysis. Next, by mapping qualitative ratings to numerical values, we captured the intended constructs of the study. This process confirms that our measurement instrument reflects only the theoretical dimensions this study measures, thereby ensuring *Construct Validity*.

While the study implements internal and construct validity, the sample size of 18 students is relatively modest: however, we analysed altogether 195 feedback points. This controlled, homogeneous group allowed for a focused assessment of the framework as a pilot study. Yet, we acknowledge that it may restrict the external validity.

6. Conclusions & Future Work

Analysing student feedback on a course and integrating the outcome into the course, not only improves the quality of teaching but also benefits students in their learning process. In this study, it is shown that the use of Generative tools like ChatGPT in evaluating student feedback to support teachers could be applied to practical workflow that could significantly lessen the workload of teachers. It can be determined from our preliminary study observations(where we analysed 390 student feedback points), that GPT-4 enhances evaluation consistency and speed, but human judgment remains vital, and thereby supports a dual-evaluator approach than a complete autonomous approach for comprehensive student feedback analysis.

In future research, we aim to construct a simple dual-evaluator approach involving both the teacher and GPT, that complements and retains the positive features of both of them. We further plan to enhance the use of GPT-4 for evaluating student feedback with larger data-sets and explore the capabilities of newer GPT reasoning models and other generative AI tools in analyzing and responding to sentiments expressed in feedback.

Acknowledgments

This work has been supported by FAST, the Finnish Software Engineering Doctoral Research Network, funded by the Ministry of Education and Culture, Finland.

Declaration on Generative AI

The authors used the GPT-4 API to evaluate student feedback as part of the methodology for this pilot study. Additionally, GPT-4 was used to debug code related to heat map generation. After using these tools, the authors reviewed the outputs for accuracy to the best of their knowledge.

References

[1] B. Marr, A short history of chatgpt: How we got to where we are today, https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/, 2023.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).

[3] OpenAI, Chat with openai, 2024. URL: https://help.openai.com/en/collections/3742473-chatgpt.

[4] C. D. Carly Steyn, A. Sambo, Eliciting student feedback for course development: the application of a qualitative course evaluation tool among business research students, Assessment & Evaluation in Higher Education 44 (2019) 11–24. doi:10.1080/02602938.2018.1466266.
[5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. arXiv:2303.12712.
[6] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, arXiv preprint arXiv:2005.00661 (2020).
[7] M. Ling, Chatgpt (feb 13 version) is a chinese room, ArXiv abs/2304.12411 (2023). doi:10.48550/arXiv.2304.12411.
[8] Y. Meron, Y. T. Araci, Artificial intelligence in design education: evaluating chatgpt as a virtual colleague for post-graduate course development, Design Science 9 (2023). doi:10.1017/dsj.2023.28.
[9] Y. Zhou, R. Tang, Z. Yao, Z. Zhu, Navigating the shortcut maze: A comprehensive analysis of shortcut learning in text classification by language models, 2024. URL: https://arxiv.org/abs/2409.17455. arXiv:2409.17455.
[10] H. Li, D. Guo, W. Fan, M. Xu, J. Huang, Y. Song, Multi-step jailbreaking privacy attacks on chatgpt, ArXiv abs/2304.05197 (2023). doi:10.48550/arXiv.2304.05197.
[11] H. Alkaissi, S. I. McFarlane, Artificial hallucinations in chatgpt: Implications in scientific writing, Cureus 15 (2023). URL: https://api.semanticscholar.org/CorpusID:257037938.
[12] T. Tricker, M. Rangecroft, P. Long, Bridging the gap: an alternative tool for course evaluation, Open Learning: The Journal of Open, Distance and e-Learning 20 (2005) 185–192.
[13] L. S. Brew, The role of student feedback in evaluating and revising a blended learning course, The Internet and Higher Education 11 (2008) 98–105. URL: https://www.sciencedirect.com/science/article/pii/S1096751608000249. doi:https://doi.org/10.1016/j.iheduc.2008.06.002.
[14] W. Dai, J. Lin, H. Jin, T. Li, Y.-S. Tsai, D. Gašević, G. Chen, Can large language models provide feedback to students? a case study on chatgpt, in: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), 2023, pp. 323–325. doi:10.1109/ICALT58122.2023.00100.
[15] W. Dai, Y. Tsai, J. Lin, A. A. Aldino, F. Jin, T. Li, D. Gašević, G. Chen, Assessing the proficiency of large language models in automatic feedback generation: An evaluation study, 2024. doi:10.35542/osf.io/s7dvy.
[16] M. Morales-Chan, H. R. Amado-Salvatierra, J. A. Medina, R. Barchino, R. Hernández-Rizzardini, A. M. Teixeira, Personalized feedback in massive open online courses: Harnessing the power of langchain and openai api, Electronics 13 (2024) 1960. doi:10.3390/electronics13101960.
[17] M. Zhang, E. Lindsay, F. B. Thorbensen, D. B. Poulsen, J. Bjerva, Leveraging large language models for actionable course evaluation, Student Feedback to Lecturers (2024). doi:10.48550/arxiv.2407.01274. arXiv:2407.01274.
[18] J. Steiss, T. Tate, S. Graham, J. Cruz, M. Hebert, J. Wang, Y. Moon, W. Tseng, M. Warschauer, C. B. Olson, Comparing the quality of human and chatgpt feedback of students' writing, Learning and Instruction 91 (2024) 101894. URL: https://www.sciencedirect.com/science/article/pii/S0959475224000215. doi:https://doi.org/10.1016/j.learninstruc.2024.101894.
[19] A. Fuller, K. A. Morbitzer, J. M. Zeeman, Exploring the use of chatgpt to analyze student course evaluation comments, BMC Medical Education 24 (2024) 423. URL: https://doi.org/10.1186/s12909-024-05316-2. doi:10.1186/s12909-024-05316-2.
[20] DeepL GmbH, Deepl translator, https://www.deepl.com/translator, 2023. Accessed: 2023-04-23.
[21] O. Weerakoon, feedback_analyser_with_chatgpt, https://gitlab.utu.fi/osweer/sunflower, 2024.
[22] O. Weerakoon, thesis_heatmaps, https://gitlab.utu.fi/osweer/thesis_heatmaps, 2024.
[23] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, A. Kocoń, B. Koptyra, W. Mieleszczenko-Kowszewicz, P. Miłkowski, M. Oleksy, M. Piasecki, Łukasz Radliński, K. Wojtasik, S. Woźniak, P. Kazienko, Chatgpt: Jack of all trades, master of none, Information Fusion 99 (2023) 101861. URL: https://www.sciencedirect.com/science/article/pii/S156625352300177X. doi:https://doi.org/10.1016/j.inffus.2023.101861.

[24] J. Rytilahti, O. Weerakoon, Ai tools for study material generation, 2023. Retrieved March 9, 2024, from https://www.utu.fi/en/university/faculty-of-technology/computing/ai-tools-for-study-material-generation.