2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
Paper ID #51273

Beyond Sentiment: [Artificially?] Intelligent Synthesis of Student Evaluation Narratives

Dr. Jennifer Retherford, The University of Tennessee, Knoxville

Dr. Retherford is an alumna of the University of Nebraska, Omaha, and received her graduate degrees from Vanderbilt University. She currently teaches a variety of courses supporting the department of Civil & Environmental Engineering at the University of

Dr. David A Saftner, University of Minnesota Duluth

David Saftner is an Associate Professor at the University of Minnesota Duluth. He received a BS in Civil Engineering from the United States Military Academy and MS and Ph.D. in Civil Engineering from the University of Michigan.

Kun Zhang, University of Minnesota Duluth

An Urban Hydrologist!

Abdullah Haroon, University of Minnesota Duluth

Beyond Sentiment: [Artificially?] Intelligent Synthesis of Student Evaluation Narratives

Abstract

Processing student end-of-semester course evaluations can be a challenging task for professors seeking to not only affirm their status as a qualified educator but also provide meaningful learning experiences for the students they influence. While professors attempt to use course evaluations for self-improvement and course development, the emotional aspects of reading student evaluations can overshadow the experience. While students are adept at understanding quality classroom engagements, they often lack training in providing quantitative feedback and experience in framing qualitative feedback in a constructive manner. The challenge of parsing through poorly written or personally directed feedback can deter faculty members from thoughtfully engaging with the comments. Faculty in various career positions may be constrained, which further disconnects the feedback from implemented change. Early career professors often need to prioritize research responsibilities, whereas mid- or late-career professors often take on additional administrative roles that divide their attention from incremental course improvements. However, most faculty do seek the opportunity to ensure the classroom they manage is positively received by students and able to serve the purpose of supporting the teaching and learning cycle. Generative artificial intelligence (AI) platforms may provide a solution to isolating constructive feedback while filtering evaluations which might be detrimental to the faculty's growth mindset. This study seeks to explore a framework for applying AI platforms towards student course evaluation synthesis leading towards improving actionable response. A study was performed to process student course evaluations for four faculty at different career stages. The team defined their own parameters for the intention of processing student feedback using AI. The paper discusses the process for developing their method and reflections made by each participant. Closing remarks articulate the value found in using AI for this purpose but also the unexpected value of developing a peer cohort.

Introduction

The purpose of this case study is to describe the effort four university instructors took to use artificial intelligence (AI) to better utilize student rating of teaching (SRT) feedback to improve their teaching. Each instructor had their own personal motivations for undertaking the project; however, common to all was the desire to get more out of the SRT feedback. The authors' intentions were to independently choose AI platforms and prompts to address their personal motivations. After refining the prompts independently, each author would use the other authors' prompts to analyze their own SRT feedback. The lessons learned are summarized below. Additionally, and unintentionally, the authors found that through discussing the results with each other, additional clarity and ideas for improvement emerged, highlighting the importance of a community of peers.

Literature demonstrates that the authors are not alone in their motivations for undertaking this project, as SRTs present significant challenges for both students and faculty. At the end of an academic term, students are often under significant stress, which can color their perceptions.

Furthermore, the high volume of SRTs they are required to complete often leads to "survey fatigue," potentially compromising the quality of their feedback [1]. Morley [2] noted that while individual students may be consistent in their own views, they often lack consistency with one another, making it difficult to find a cohesive narrative in the feedback. For instructors, the difficulty lies in the emotional and technical burden of interpreting this feedback. Lakeman et al. [3] reported that 81% of faculty responding to an anonymous survey reported stress from their SRT feedback. This stress is compounded by a widespread lack of training in data interpretation. As reported by Linse [4], faculty frequently overinterpret small, statistically insignificant differences between numerical SRT results from semester to semester. In addition, the data itself is inherently "noisy;" Curby et al. [5] demonstrated that 62% of score variance stems from situational factors rather than actual teaching quality. For early-career faculty, this inability to distinguish signal from noise is particularly distressing, as the high stakes of tenure evaluations often force them to prioritize moving these numerical needles over fostering genuine student engagement [6]

The authors' literature review did not find many examples of recommendations for SRT interpretation. Royal [7] proposed a guide for faculty and administrators to assess the interpretative validity of SRT scores. Meyer et al. [8] developed a methodology, using item response theory and Wright mapping methods, to interpret SRT scores and identify ways to refine teaching. Neither of these studies addressed the authors' motivations for undertaking this project. Finally, before the development of the most recent generative AI platforms such as ChatGPT, Gemini, Grok, and Copilot, there have been some studies using various natural language processing, machine learning, and deep learning tools to process SRTs, but their functionalities remain at sentiment classification, prediction, and statistical and mathematical analysis [9-11]. The authors found that no study has assessed the effectiveness of the most recent generative AI platforms in interpreting SRTs.

Method

The team brain-stormed a methodology with the intention of retaining a high level of individualism in terms of both tools and intentions. Common to all participants' motivation was to determine the value of using AI to interpret SRTs to inform future course improvements. The methodology was developed through a series of co-author meetings which were collaborative and open-ended. As a result of these meetings, the components of the method can be discussed in terms of boundaries of our exploration, agreement on elements of AI use, and compilation of relatable themes.

First, co-authors identified their motivations for exploring the posed research question, explored use of AI with their SRTs without any peer influence. The team then shared their experiences and refined their prompts, shared their refined prompts, and then used the peer prompts to study their SRTs, and again shared reflections. After initial exploration, the team recognized the need to define the boundaries of data and restrict their analysis to one AI platform. The team agreed to reduce the study to only one academic year and one course. The study's intention was to prioritize developing a system for evaluating SRTs and therefore more data such as inclusion of additional courses or additional semesters was considered unnecessary.

The authors agreed to maximize the opportunity for AI design by placing no restrictions on platforms or prompts, recognizing different institutions' AI protocols, security and privacy policies, as well as preferences of those attempting to replicate this process, which could include campus-encouraged platforms but could extend to personal paid subscriptions. Each author independently designed their own prompts. No constraints were placed on any individual, allowing for flexibility in developing one prompt, a series of prompts, evaluating semesters independently or combined, and in using or ignoring AI-generated suggestions for continuing the chat threads. Each participant was asked to develop their prompts using their personalized SRTs. The SRTs were already clear of student identifiers as standard protocol at the participants' institutions. After using their own prompts and student feedback, the participants then provided the prompts to the group to allow others to use their same phrasings.

After individual prompt design was achieved, the team was tasked with using the prompts from the other co-authors in the preferred platform. Individuals were asked to reflect on the combined process of identifying their motivation, designing their prompts, and using the prompts of their peers. Ultimately, because of the team's regular meetings, the merit of using different prompts was considered advantageous and likely more beneficial than developing a singular AI prompt.

Evaluation & Discussion

AI was used in a collaborative effort to process student evaluations and investigate the merit of AI as a tool to improve actions responding to course feedback. The team of four co-authors were deliberately recruited to represent different stages of career progression and career paths (see Table 1). To make the process approachable and flexible, each participant defined their own plan of study (see Table 2). Choices of the courses to process and the AI platform were made based on preferences but also availability as constrained by university AI policies. For some harmony, the team did agree to review only feedback received in the 2024-2025 academic year. The team agreed to use the most recent term (Spring 2025), which was considered to provide sufficient content for the experiment. Student evaluations from a single year kept the experimental effort to an agreed upon manageable level and considered to minimize challenges related to big or sparse data post-processing. Each member of the cohort applied the process (see Appendix for their unique AI prompts) and reflected on their experiences, considering these specific prompts:

1. What was your personal motivation for using AI to process student feedback?
2. What did you learn from this exercise?
3. What did you learn from other people's prompts?
4. What did you learn from the community experience?

TABLE 1: Participant Demographics

Label	Name	Years in Academia	Career Path	Career Stage
1	Dr. Saftner	15 years	Tenure Track	Tenured
2	Dr. Retherford	13 years	Non-tenure Track	Reached highest promotion level
3	Dr. Zhang	3 years	Tenure Track	2 years until tenure
4	Dr. Haroon	3 years	Tenure Track	2 years until tenure

TABLE 2: Participant Preferences & Data Choices

Label	Name	AI Platform	Course Processed	Typical Student Standing
1	Dr. Saftner	Google Gemini 3	Soil Mechanics	Undergraduate Juniors
2	Dr. Retherford	Microsoft Co-Pilot	Senior Design Project, 2nd Term	Undergraduate Seniors
3	Dr. Zhang	Grok 4.1 (X.AI Corp.)	Fluid Mechanics	Undergraduate Sophomores
4	Dr. Haroon	Google Gemini Pro	Design of Reinforced Concrete	Undergraduate Juniors, Seniors

Professor 1

Personal Motivation

Early in my career, SRT feedback provided impactful suggestions for improvement. As I improved my courses, I struggled to find actionable feedback, often thinking "that's what they say about me." Conflicting feedback (i.e., too fast vs. too slow, too much homework vs. frequent homework helped) also complicated my personal analysis. I previously used SRT feedback to improve more often than I have recently. I want to get more out of my SRT feedback. Therefore, I tried to answer this question. Can AI help me find improvements that I have been missing?

Reflection

My initial attempts at prompts were not ideal because I did not restrict the length of the AI response and received a regurgitation of my students' comments. Additionally, I learned that if I did not include both comments focused on things I should sustain, as well as improve, AI did not account for conflicting feedback (i.e., too much homework vs. frequent homework helped). The prompt that I settled on was "based on the following student feedback on a college course, provide three points I should sustain and three points I should improve."

I repeated that prompt using SRT comments from my required, junior-level, geotechnical engineering course with a laboratory component from one through six semesters. Regardless of the number of semesters of feedback I included, the "sustain" comments were stable and reflected what I had expected from my own reading of the comments. The "improve" comments were helpful, clarifying two consistent issues regardless of the number of semester feedback I included. AI provided a third "improve" comment that varied between two primary comments depending on the number of semesters in my analysis. My biggest lesson learned was that AI was helpful in separating and clarifying related issues. For example, I missed consistent student feedback to provide more details in notes (i.e. pointing to where to find greater detail in the text), confusing it with positive student feedback on note clarity and organization. After discussing the "improve" comments with my co-authors, I developed a plan I am excited to already be implementing to address SRT feedback.

The other authors' prompts asked AI to make recommendations where I only asked AI to summarize my feedback. My learning was therefore focused on AI's recommendations. With few exceptions, I preferred the ideas our team generated on how to address issues than AI's recommendations. I was pleasantly surprised with the exceptions and will implement those improvements.

AI's feedback also reminded me of Linse's [4] and Curby et al.'s [5] work. For example, AI seemed to highlight issues that were unique to a group of students rather than my teaching in general. An example comparing Fall 2024 and Spring 2025 feedback credited me with "fixing" some issues that I put no effort into and identifying other "new issues" that were in fact consistent between the classes. Initially, I was underwhelmed by AI's recommendations. The community was more impactful than AI in helping me identify how I could continue to use my SRT feedback to improve my teaching. Following my initial prompt to simply categorize the responses, the team's brainstorming about how to best address the issues was incredibly helpful.

Additionally, my initial concerns about AI's ability to provide meaningful recommendations were lessened through discussion with the community. Our community's discussion and previous experience working with other communities helped me generate ideas to address my SRT feedback. Those with less teaching experience and, more importantly, less established support groups more likely to find AI's suggestions a helpful place to start because they are less likely to have experience and support to inform alternative suggestions. Overall, my two key takeaways from this project is that a) AI helped me organize and clarify my SRT responses and b) community, this and others developed through, among others, American Society of Engineering Education (ASEE), American Society of Civil Engineer's Excellence in Civil Engineering Education (ExCEEd) teaching workshop, my University, College, and Department, gave me the tools to best address the feedback.

Professor 2

Personal Motivation

When reviewing student evaluations, I struggle with managing my emotional response to the comments. In the past few years, I have not been able to successfully enhance my course because of this emotional stumbling block. I am interested in finding ways to more fully accept student comments and use them in productive ways to change my courses. In short, I have been looking forward to knowing: can AI be used to decouple my emotional response from my interest in making course improvements? Can AI help me prioritize my course improvements and reduce my emotional attachment to my course materials?

Reflection

Use of Microsoft's Co-Pilot did alter my emotional reaction to processing student course evaluations. While my emotional response was still high, I felt more goal-oriented in the process of allowing AI to assist me in using the responses for course improvement. Reading student feedback directly hindered my ability to see the opportunity to receive the feedback as constructive, but interacting with the AI platform provided a type of blinding so I was no longer in a dialogue with student comments but instead with a neutral party.

Prompt evolution was a valuable process, and I believe the development of initial prompts which lead to new prompts both initiated by me as well as those presented by AI allowed me to think about my real intentions in processing the student evaluations. The effort to decide my question for AI and then the series of prompts and responses improved focus on moving into my next course offering. In the past, my emotional response to student feedback tended to result more in reflection and dwelling in the past, but using AI required me to put words to the questions and I was more focused on a productive intention towards action.

Prompts designed by peers offered additional value to the process of taking student feedback and developing a plan for course improvements. The prompts designed by my three colleagues offered different lenses to evaluate the students' comments and encouraged ideas I had not considered myself. The prompts varied in complexity, which I believe generated a greater output from Co-Pilot, even though my peers were using different platforms. The use of their prompts in Co-Pilot reinforced some common themes which emerged from all prompts but also presented different perspectives. Considering my emotional response, I was positively supported by Kun's prompt designs which requested the AI platform to isolate positive themes. Abdullah's inclusion of both "Do's" as well as "Don'ts" also resulted in positive feedback which improved my reception of the Co-Pilot summaries.

AI processing of my student evaluations improved my experience, but the impact of community offered a more substantial impact on my engagement with the student feedback. The collaborative effort to discuss our goals and share our prompts opened my mind to the variety of ways we can use student evaluation comments. My emotional response to processing feedback may have been driven by a narrow series of questions I attempted to answer using the SRT comments. Peer discussions created a timeline to intentionally process my student feedback while the responsibility of reporting my progress improved my truthful accountability. Most notably, the use of other peer prompts influenced my experience, and I found the entire process more enjoyable. I feel more motivated to modify my course because I am better supported by the positive aspects I should retain and better guided by the AI-generated feedback. The AI-generated feedback has been organized better than my previous attempts. I attribute this to the external conversation between the agent and myself which required articulation of ideas and resolution towards closing action items.

Professor 3

Personal Motivation

As an early-career faculty and non-native speaker, I do not have enough confidence that I can effectively understand the hidden information within nuanced student feedback and translate those interpretations into effective pedagogical action items. Therefore, I tried to answer this question: can AI assist early-career faculty like me to effectively interpret student feedback and translate into actionable pedagogical steps?

Reflection

I selected Grok (X.AI Corp.) for the evaluation because the more widely used platforms do not fit this case study - the free version of ChatGPT does not allow me to upload more attachments, and Gemini cannot identify certain pages of my SRTs. I performed a single-turn execution with the SRT(s) attached. The prompts were developed to address my main curiosities and concerns on the SRTs, including 1) how did the student perceive my teaching, 2) what possibly caused their sentiment, 3) did I address the issues between the semesters before pursuing help from AI, and 4) what are the remaining action items to address the issues. I asked AI to analyze the possible reasons behind the feedback because I was curious and believe those analyses might enhance AI's deep thinking and potentially adjust the suggestions.

I first used my own prompt and repeated the process using other faculty's prompts. AI demonstrated a sophisticated ability to move from raw data to pedagogical theory. Across all four prompt styles, AI consistently identified the main strengths and issues reflected in the SRTs. The core strengths include exceptional clarity in explanations, being approachable and supportive, and a highly valued two-submission homework system. The primary issues include a persistent mismatch between lecture slides and the printed note packet, and a significant "misfire" regarding a new cross-class group paper project in Spring 2025. Furthermore, AI did not just list what students said; it explained why they said it and how to fix it, with the depth of the "why" and "how" differing between using different prompts. For instance, AI noted the mismatch between slides and note pack created "cognitive friction" in mathematically dense courses like Fluid Mechanics. AI further provided specific solutions, such as "*sticking more closely to the notes' order or updating the notes/slides to better match the live lecture flow*", or a "*pre-semester audit process.*"

Using different prompts does trigger varying responses from AI. I found the differences mainly lie in the level of depth in explaining "why" and "how to fix." In terms of "why," AI's response to my prompt presents the most sophisticated interpretation and analysis because my prompt explicitly asks for an "*interpretation & analysis*" of "*potential underlying causes or student perceptions*"; Others' prompts used more passive or surface-level verbs in terms of analysis, so AI either only mentioned "why" briefly or summarized the problems without explaining the causes at all. In terms of "how to fix," AI's responses vary in the actionability of the solutions provided, which was directly proportional to how explicitly the prompt asked for "concrete" or "actionable" steps. My, as well as Professor 2's, prompts used high-constraint commands, demanding "concrete" or "actionable" steps. Thus, the AI provided operational tactics. Professor 1's prompt used a more general, open-ended structure: "*... provide three points I should sustain and three points I should improve*". So, the AI provided pedagogical goals rather than technical steps. The specificity of Professor 4's prompt was between mine/Professor 2's and Professor 1's because it asked for action items but did not highlight the actionability of the solutions. Depending on your goals, I recommend refining prompts to focus on your specific priorities, i.e., emphasizing a diagnosis, an explanation, or solutions.

Having a faculty cohort to brainstorm and evaluate SRTs using AI is helpful as this process allows me to see the same data through different lenses. Through the meetings and discussions, particularly after an unsatisfactory semester, I realized that the emotional aspects of reading student evaluations also trouble me, and it can be helpful if the AI can filter out the personally directed information. Professor 4's prompt design is inspiring as he specifically asked AI to ignore certain feedback that is related to an intentional class design. This inspired me to tailor my prompt to bypass expected feedback that I may not want to address or am not ready to address yet. In addition, by talking with more senior faculty like Professor 1 and Professor 2, I also better understood the headache of mid-career professors in further improving their teaching - getting to a relatively stationary state after trying all different approaches. All of this can help me better customize my prompts for AI evaluation and tailor my teaching approaches.

Professor 4

Personal Motivation

Reading through individual student comments on SRTs and identifying constructive feedback can be time-consuming. In a large class with diverse opinions, it is not always easy to see the big-picture trends of what is working and what is not. AI can help efficiently analyze SRTs, highlight recurring student concerns, and filter out personal or non-productive comments. Therefore, I tried to answer this question: can AI-driven analysis of SRTs reveal actionable patterns of teaching practices that consistently help or hinder student learning?

Reflection

After initially playing around with the same prompts in ChatGPT and Google Gemini, I opted for the Gemini to finalize my SRT evaluation as it seemed to follow the instruction better as compared to more verbose responses from GPT. My main motivation of this exercise was to draw the big picture patterns in my student evaluation which are often overlooked when you are focusing on individual student comments and trying to figure out which aspects of the feedback require immediate attention or which of these would be the most helpful to student learning.

My first attempt at the analysis was along the lines of "here is my SRT, provide me three do's and three don'ts," which I soon realized that it is very open and given the non-deterministic nature of the AI models, the model responses would vary significantly between different platforms as well as between multiple runs of the same platform. So, there was a need to draw a box around the response. I then refined my prompts into three separate styles each focused on a specific need of information (see appendix to refer to these prompts). Prompt #1 was still quite open with a loose string of tying the suggested feedback on 'active learning' strategies. Prompt #2 focused on isolating "noise". Class aspects that were intentional but could be perceived by students as hindrance, were identified in the prompt to avoid recommendations around these topics. Finally, Prompt #3 focused on longitudinal analysis of student feedback by comparing SRTs from two consecutive semesters to identify any long standing and recurring issues that are consistent, issues that seems to be resolved, or new issues that arose in the latest class delivery.

The AI showed a strong capability to synthesize qualitative student feedback into clear, actionable items. Across the three prompt styles, the AI consistently identified the "Structured Note Packets" as the cornerstone of the course's success, with students in both semesters describing them as "phenomenal" and "efficient." On the other hand, the primary areas for improvement were consistently identified as a need for more hands-on exercises to visualize structural behavior and a delivery style that occasionally drifted into monotony.

A key finding in this exercise was the AI's ability to filter "noise" based on instructor-defined parameters. In Prompt #1, without context, the AI flagged variable notation as a "Don't" as it confuses students. However, once the prompts (Prompt #2 and #3) explicitly stated that the notation complexity reflects state-of-practice and homework disconnect was an intentional pedagogical choice to promote independent problem solving, the AI successfully ignored those complaints. Instead, it diverted its analysis to focus on delivery, approachability, and specific technical gaps (such as the lack of diverse examples). It provided feedback that was far more aligned with the instructor's specific teaching philosophy.

Another thing of note is that the structure of the prompts significantly altered the nature of the advice. Below is a description of the subtle difference in the model responses guided by the change in the prompt requests:

- **Prompt #1 (Theory-Based):** By asking to tie suggestions to "active learning strategies," the AI provided a high-level pedagogical framework, utilizing terms like "scaffolding" and "fostering psychological safety." This offered a theoretical "why" behind the SRT comments.
- **Prompt #2 (Constraint-Based):** By adding specific constraints ("ignore X, Y, and Z"), the AI moved away from theory and provided highly tactical "dos and don'ts" focused on course administration and immediate classroom behavior (e.g., "don't assume silence implies understanding").
- **Prompt #3 (Longitudinal Analysis):** This prompt revealed the AI's capacity for trend tracking. It successfully identified that the "timeliness of homework posting" issue from Fall 2024 was resolved in Spring 2025 but flagged a new issue regarding "approachability" that emerged only in the second semester.

A comparison of the model responses using my own prompts vs prompts generated by other faculty revealed that my own prompts produced results that were much more specific and focused. The prompts from other faculty were open ended when I posed them against my research question. This broad nature of the other prompts caused the analysis to focus on common student complaints - such as the perceived disconnect between homework and examples or the relevance of the textbook. These standard observations were not helpful because those aspects of the course are intentional. In contrast, my Prompt 2 successfully filtered out these issues, allowing the analysis to uncover deeper actionable items that would otherwise be hidden, such as the specific student request for a cantilever example and feedback regarding "monotonous" lectures. Additionally, my Prompt 1 proved effective by explicitly requesting "active learning" strategies, which yielded professional pedagogical recommendations rather than the generic engagement advice generated by the other prompts.

This exercise highlights that AI is most effective when the instructor acts as a gatekeeper of context. By explicitly telling the AI what to ignore (intentional friction) and asking it to compare data over time, the tool shifts from a simple summarizer to a strategic partner in course improvement. Finally, although I prefer my own prompts over others as they are most in sync with what I am trying to extract from my SRT, this should not lead anyone away from the importance of having a community around this exercise. Engaging with colleagues allows them to act as essential sounding boards, bringing fresh insights that challenge my own assumptions and reveal blind spots I might otherwise miss. By sharing our diverse strategies and prompt structures, we can learn from one another's approaches, discovering new ways to interpret data that we wouldn't have found in isolation, and turn a solitary evaluation process into a collective effort for pedagogical improvement.

Closing Remarks

A flexible method for processing SRT feedback using AI was developed and performed with positive results. The initial goal of the process was to draft a framework for using AI in a systematic and repeatable way which could be adopted by another faculty. Through brainstorming, a strict regimen was discouraged and instead a flexible approach was developed. The approach allowed for enhanced personal ownership of the process and minimized distractions associated with platform advantages, disadvantages, and unrelated data processing hurdles. Three takeaway themes were identified:

- Working in a cohort enhanced prompt development for most individuals in the group.
- Working in a flexible system allows quick adoption and fosters a clearer sense of ownership in the process.
- Use of AI to process SRTs was met with mixed opinions. Each participant reacted differently to the value of using AI as a tool. Each participant found varying aspects of AI's analysis more helpful than others; however, all agree the AI tool has utility in interpreting SRT results.

Ultimately, the process of developing an agreed upon method, engaging in the method, and intermittent interaction with peers was identified as the recommended practice for processing student feedback. AI tools did reduce the time most participants used to read and analyze their SRT comments, allowing for more time to really design the questions they were seeking to address. The effort of prompt design and responding to AI output partnered well as a common baseline for small group discussion. Review of other people's prompts and discussion of those prompts improved the participants' ability to isolate the real outcomes they were defining for their personal motivation. Each faculty member naturally criticized the AI feedback and meetings with the cohort of peers helped in determining which AI output had merit and which deserved further review.

References

[1] N. Fjortoft, "A reflection of faculty and course evaluations," American Journal of Pharmaceutical Education, vol. 79, no. 9, p. 129, 2015.

[2] D. Morley, "Assessing the reliability of student evaluations of teaching: choosing the right coefficient," Assessment & Evaluation in Higher Education, vol. 39, no. 2, pp. 127–139, 2014.

[3] R. Lakeman, R. Coutts, M. Hutchinson, D. Massey, D. Nasrawi, J. Fielden, and M. Lee, "Stress, distress, disorder and coping: The impact of anonymous student evaluation of teaching

on the health of higher education teachers," Assessment & Evaluation in Higher Education, vol. 47, no. 8, pp. 1489–1500, 2022.

[4] A. R. Linse, "Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees," Studies in Educational Evaluation, vol. 54, pp. 94–106, 2017.

[5] T. Curby, P. McKnight, L. Alexander, and S. Erchov, "Sources of variance in end-of-course student evaluations," Assessment & Evaluation in Higher Education, vol. 45, no. 1, pp. 44–53, 2020.

[6] S. Acker and M. Webber, "Made to measure: Early career academics in the Canadian university workplace," in Academic Life in the Measured University, Routledge, 2020, pp. 89–102.

[7] K. Royal, "A guide for making valid interpretations of student evaluation of teaching (SET) results," Journal of Veterinary Medical Education, vol. 44, no. 2, pp. 316–322, 2017.

[8] J. P. Meyer, J. B. Doromal, X. Wei, and S. Zhu, "A criterion-referenced approach to student ratings of instruction," Research in Higher Education, vol. 58, no. 5, pp. 545–567, 2017.

[9] C. Dann, P. Redmond, M. Fanshawe, A. Brown, S. Getenet, T. Shaik, X. Tao, L. Galligan, and Y. Li, "Making sense of student feedback and engagement using artificial intelligence," Australasian Journal of Educational Technology, vol. 40, no. 3, pp. 58–76, 2024.

[10] A. S. Sunar and M. S. Khalid, "Natural language processing of student's feedback to instructors: A systematic review," IEEE Transactions on Learning Technologies, vol. 17, pp. 741–753, 2023.

[11] M. J. Parker, C. Anderson, C. Stone, and Y. Oh, "A large language model approach to educational survey feedback analysis," International Journal of Artificial Intelligence in Education, vol. 35, no. 2, pp. 444–481, 2025.

Appendix: Personal AI Prompts

Dave

Based on this student feedback on a college course [I pasted in my student comments below], provide three points I should sustain and three points I should improve.

Jenny

1. Hello! I am looking for your assistance in identifying actionable themes from some student feedback I received. Can you please review this text? Can you please return to me a summary which meets these goals? 1. please remove any personally directed feedback so I can better receive the input without feeling emotional about their opinion of me as a person. 2. please isolate 1-3 themes which appear to develop. 3. please recommend action items you think I can take to improve based on any requests from the themes. Here are the reviews:
2. {posted the responses from one open-ended question from my evaluations}
3. Thanks. I don't need a visual or reflection document. Can you follow the same three requests for this different prompt given to my students?
4. {posted the responses from a second open-ended question from my evaluations}
5. yes - I would appreciate a consolidated and unified summary.
6. Here are responses to the same two questions from a previous semester. Can you please tell me if there are any differences between these responses and the consolidated summary you created? Responses to the first prompt: {posted responses…to the second prompt…responses}
7. yes - I'd like to see a side-by-side comparison table
8. Lastly – of the recommended action items you provided me, can you rank them based on efficiency? Can you list the action items in these three ways: (1) potential to have the most positive impact on my students (2) lowest effort on my part (3) a composite which balances best impact on my students for moderate effort on my part
9. yes - can you do both? Create an action prioritization matrix and create a visual chart showing impact vs. effort? Thanks!

Kun

1. I am a university instructor and I have received end-of-semester student feedback from my class as attached. My goal is to understand the core themes and generate a plan for improvement. Please analyze the feedback provided and give me three distinct outputs:
   #1. Key Takeaway: Provide a concise, overarching key takeaway that summarizes the overall sentiment and main finding. Identify and list the 4-6 most prevalent and significant themes (2-3 positive; 2-3 negative) that appear throughout the feedback.
   #2. Interpretation & Analysis: For the most critical negative theme, provide a deeper analysis of the potential underlying causes.
   #3. Concrete Action Items: Based on your analysis, suggest 2-3 high-priority, specific, and actionable steps I can take to improve the course/instruction for the next semester, starting with the low-hanging fruits. Ensure these steps are concrete and address the identified negative themes directly.
2. The attached file is the student review feedback in the following semester for the same course. Please do the same analysis and give me three distinct outputs, including #1 Key Takeaway, #2 Interpretation & Analysis, and #3 Concrete Action Items.
3. Please compare the two feedbacks provided and give me three distinct outputs:

#1. Provide a concise, overarching key takeaway that summarizes the overall trajectory of student sentiment and learning outcomes, and quantify the improvement.
#2. Analyze the shifts in student feedback by listing the following four categories of themes:

- Enduring Strengths: Positive themes or high-scoring metrics present in both semesters.
- Lost Strengths: Positive themes from the first semester that disappeared or saw a measurable decline in the following semester.
- Resolved Issues: Negative themes from the first semester that were successfully improved in the following semester.
- Persistent Issues: Negative themes that remain problematic in the following semester.

#3. Concrete Action Items: Based on your analysis, suggest 2-3 things I should keep doing to maintain the strengths and 2-3 high-priority, specific, and actionable steps I can take to pick up the lost strengths and resolve the persistent issues for the next semester, starting with the low-hanging fruits.

Abdullah

1. Here is a PDF of my end-of-the-semester course evaluation by students in my Design of Reinforced Concrete class. Evaluate the student evaluation and respond with three action items for dos and don’ts that would help me improve my teaching. Be brief and tie your suggestions to active learning strategies.
2. Here is a PDF of my end-of-the-semester course evaluation by students in my Design of Reinforced Concrete class. Throughout the course, I remind them that the perceived disconnect between in-class examples and homework problems is by design to force the ‘engineer’ out of them. Their prerequisite coursework for this class acts as the bridge between this gap. Additionally, there is no required textbook for this class; the ACI 318-19 is an optional buy and highly recommended for structure-focused students. Finally, the perceived repetition of variables in some design notations is a requirement of trade and how the design codes are written. Ignore any student concerns related to these issues. Respond with three action items for dos and don’ts that would help me improve my teaching. Be brief and limit your response to the contents of the PDF only.
3. Here are the PDFs of my end-of-the-semester course evaluation for the Design of Reinforced Concrete class for Fall 24 and Spring 25 semesters. Throughout the course, I remind them that the perceived disconnect between in-class examples and homework problems is by design to force the ‘engineer’ out of them. Their prerequisite coursework for this class acts as the bridge between this gap. Additionally, there is no required textbook for this class; the ACI 318-19 is an optional buy and highly recommended for structure-focused students. Finally, the perceived repetition of variables in some design notations is a requirement of the trade and how the design codes are written. Ignore any student concerns related to these issues. Evaluate the student feedback and do the following:
    a. Identify key issues that are common in both semesters.
    b. Identify issues that were present in Fall 24 but seem resolved in Spring 25.
    c. Identify any new student consent that was not present in Fall 24 but showed up in Spring 25.
    d. Identify praises that are common in both semesters.

Provide a concise bulleted list for each point.