

Leveraging Large Language Models for Scalable Analysis of the End-of-the-Course Student Feedback

Jelena Jovanovic¹, Irena Vodenska², and Vladan Devedzic^{3,1,4}

¹ Faculty of Organisational Sciences, University of Belgrade

Jove Ilica 154, 11000 Belgrade, Serbia

jelena.jovanovic@fon.bg.ac.rs, vladan.devedzic@fon.bg.ac.rs (corresponding author)

² Administrative Sciences Department, Metropolitan College, Boston University

1010 Commonwealth Avenue, Boston, MA 02215

vodenska@bu.edu

³ Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

⁴ Serbian Academy of Sciences and Arts, Knez Mihailova 35, Belgrade, Serbia

Abstract. Analyzing open-ended student feedback in course evaluations is a labor-intensive task due to the unstructured and complex nature of natural language. While Large Language Models (LLMs) offer significant potential for automation, a well-defined methodology for their application in analyzing student feedback remains underdeveloped. This paper addresses this gap by proposing an LLM-based feedback analytics pipeline designed to transform students' open-ended feedback into structured, actionable insights. The pipeline consists of three sequential stages: (i) segmenting student feedback into semantic units and assigning polarity (sentiment) to those units; (ii) topical classification of semantic units, and (iii) summarization of units within each topical category. By systematizing these processes, the proposed method enables educators and course managers to efficiently derive meaningful patterns from vast datasets of student opinions. We evaluated the proposed method using a comprehensive dataset from several editions of a U.S. university course, yielding encouraging results of the method's effectiveness. This research provides a scalable, generic methodology for (semi-)automated feedback analysis, ultimately supporting data-informed improvements in teaching and course management.

Keywords: Student Feedback, Course Evaluation, Natural Language Processing, Large Language Models, Qualitative Coding.

1. Introduction

Student evaluations of teaching provide crucial quantitative and qualitative data for enhancing course design and delivery in higher education. This data can be analyzed and used by teachers and school management to improve course¹ organization, update course contents and resources, enhance course assignments and grading, and refine teachers' approaches and their communication with students [24]. Moreover, the data collected from

¹ Note that throughout the paper we use the term course to refer to an individual subject, not a whole degree

student evaluations may serve to institutional management as an important source of indicators of teachers' contributions for faculty promotion and tenure prospects.

It is rather straightforward to automate the analysis of quantitative data from student evaluations. However, it is much more challenging to analyze qualitative data, since it originates from open-ended student responses, which, as any textual data, needs to go through a natural language processing pipeline before it can be effectively used for obtaining insights from student comments. This paper presents a methodology based on Large Language Models (LLMs), aimed at facilitating analysis of this kind of qualitative data, ultimately leading to relevant information for instructors, course coordinators, and other relevant stakeholders in higher education institutions.

A key concept in the proposed methodology is qualitative coding, the process of categorizing (annotating) segments of text (in this case open-ended student responses), i.e. identifying major themes in these segments [36]. Traditionally, this has been a manual, labour-intensive task and as such significantly impeded large-scale qualitative data analysis. Aiming to overcome this challenge, researchers are increasingly using LLMs in qualitative coding of text in various educational applications, and the results are intriguing and encouraging. For example, in their 2024 study, Hou et al. [20] suggested fine-tuning of GPT models for better accuracy in coding along context-dependent dimensions of text (i.e., theorizing, integration, reflection), whereas the accuracy was satisfactory in coding along context-independent dimensions (i.e., appraisal, questioning, social, curiosity, surprise). Zhang et al. [47] explored the performance of their GPT-3-based tool in both inductive coding (discovering major themes in text segments) and deductive coding (labeling text segments according to a pre-specified coding scheme) on simulated and real datasets. Using inter-rater reliability measures, they found substantial agreement between human coders and their tool in various coding scenarios. Barany and colleagues used GPT-4 in an attempt to automate the codebook development process [3]. They concluded that LLMs could be valuable for improving qualitative codebooks, but human involvement in their development remained essential. Liu et al. [27] used GPT-4 in the context of qualitative coding of semi-personalized tutoring session transcripts and student observations and behavior across three different courses. They have specifically studied the inter-rater agreement between human coders and GPT-4 when using prompting techniques like Zero-shot, Few-shot, and Few-shot with contextual information. The results indicate that for good results each technique should be tailored to the specific properties of the construct and context being analyzed.

Yet, research related to using LLMs to support students' qualitative evaluations of teaching and courses remains scarce. These evaluations most often come in the form of natural-language answers/comments to open-ended questions, where challenges include: (i) the volume and unstructured nature of open-ended comments, and (ii) countless ambiguities in such comments, originating in the commenter's writing style and language use, thus exacerbating the process of "making sense" out of the overall student feedback. A few researchers have taken the challenges. For instance, Peña-Torres [35] conducted sentiment analysis of student evaluations using an LLM fine-tuned with a crowdsourced dataset of students' comments and reported 93% accuracy in the classification of the comments. Gul et al. [18] compared multi-task learning with distinct LLMs on the tasks of aspect extraction and sentiment analysis across more than 11,000 comments in course evaluations from a public university. Maram et al. [30] took a different approach – they

developed three agent-based conversational LLM tools, one for teachers to create and test prompts to gather student feedback on the course and teaching, one for students to provide feedback, and one for teachers to view and analyze this feedback. More closely related to the current paper is the recent work of Parker and colleagues [34], which demonstrates the capability of LLMs to streamline distinct natural language processing workflows relevant to the analysis of students' responses to open-ended course evaluation questions. Similarly, Fuller et al. [17] examined the efficiency and effectiveness of LLMs in identifying the main themes in students' open-ended responses to a course evaluation survey.

Having conducted a detailed analysis of these and similar efforts (see Section 2.2), we have identified a research gap that has driven the current study: the lack of a clearly defined and fairly general LLM-based methodology to automate the processing and interpretation of student open-ended feedback on end-of-the-course surveys. Thus this study proposes such a method, with the ultimate goal of enabling the transformation of raw, qualitative feedback into structured, actionable insights for instructors and course administrators. We refer to this method as *feedback analytics pipeline* and demonstrate it on a rich dataset of student end-of-the-course feedback, comprising more than a decade of feedback data from a course taught at a U.S. university.

The following sections detail the relevant literature review, the proposed feedback analytics pipeline, the application and evaluation of the pipeline on the aforementioned dataset from a U.S. university, and a discussion of the findings and future work.

2. Literature Review

2.1. End of the course student feedback

Student feedback is a central data source in educational research and scholarship of teaching and learning, providing relevant information about learners' experiences with instruction, assessment, and the learning environment. In higher education, this evidence most commonly takes the form of student evaluations of teaching (SETs), typically combining Likert-like ratings with open-ended comments. Historically, quantitative evaluations have been the focus of analysis due to their ease of evaluation, while the (rich) qualitative student feedback has been largely overlooked, given the complexity of manual coding. However, the qualitative component of student feedback is important for providing enhanced signals of course and instructor quality, content, pedagogical approaches, and ultimately, for course improvement.

Research positions student feedback as one component of an evidence-informed approach to teaching evaluation and course enhancement, ideally triangulated with peer review, teaching artifacts, and indicators of student learning [5, 31, 40]. Conceptually, student feedback can be framed in two complementary ways. First, as *measurement*, a set of instruments intended to capture distinct dimensions of instructional quality (e.g., organization, clarity, engagement). Second, as *student voice*, a means of incorporating learners as legitimate informants and partners in improving teaching and curricula [7, 14]. This framing matters because it shapes both the purposes of feedback (formative improvement vs. summative judgment) and the analytic methods used to interpret it. Both perspectives are important to our study as our objective is to support continuous improvement of course offerings and instruction through the integration of students' voice (i.e. feedback).

Open-ended student comments can add diagnostic value beyond numeric ratings by providing context, identifying specific issues with the course design, and offering actionable suggestions. Empirical work on written comments shows that students frequently address topics such as clarity of instruction, course organization, relevance/interest, assessment demands, and instructor-student relations [9,33]. Written feedback is often particularly useful for identifying *what* to change and *why* a rating might be high or low, thereby supporting targeted improvement. Moreover, qualitative feedback aligns with an improvement-oriented view of teaching as iterative design. In constructive alignment and reflective teaching traditions, instructors refine intended learning outcomes, teaching/learning activities, and assessment through cycles of evidence and reflection [6, 10]. Importantly, mid-semester feedback processes (e.g., small group instructional diagnosis or structured midterm surveys) can be especially useful as they enable changes while a course is still running [13, 15].

Because open-ended student comments can be voluminous, idiosyncratic, and uneven in quality, their value depends heavily on systematic analysis. Rigorous approaches typically involve qualitative coding, development of a codebook, iterative refinement, and checks on credibility (e.g., intercoder agreement for descriptive coding, audit trails, and reflexive memoing) [8,36,38]. When conducted carefully, qualitative analysis can (i) separate actionable course-design issues from non-actionable preferences, (ii) detect recurring patterns across cohorts, and (iii) surface equity-relevant experiences that may be obscured in averages. The best practice is to integrate qualitative themes with quantitative indicators. For example, instructors and departments can map themes (e.g., “assessment clarity”) to course elements (rubrics, workload pacing, alignment) and then evaluate whether subsequent iterations show changes in both student feedback and learning outcomes [5,6].

Recent work explores computational approaches to analyzing large corpora of student comments, including sentiment analysis, topic modeling, and structural topic modeling. Topic modeling using LDA (Latent Dirichlet Allocation) has been proposed as a scalable alternative to manual coding, enabling institutions to identify common themes and examine variation by course level or subject area [21,42]. Structural topic modeling extends this by incorporating covariates (e.g., discipline, enrollment, instructor characteristics) and can link comment themes to numeric SET outcomes [49]. These approaches can support system-level sensemaking, but they still require validation against human interpretation and careful governance to avoid reifying bias [19,41].

Overall, the literature supports a balanced conclusion: student feedback is most valuable when treated comprehensively as *evidence for improvement* rather than a single high-stakes score, and when qualitative comments are analyzed systematically using transparent, methodologically sound approaches that enhance actionability and fairness [5, 19, 40].

2.2. LLMs in the analysis of educational content and feedback

Open-ended text from course evaluations and other engagement surveys and narratives, contains high-value signals that are often underutilized because key stages of traditional qualitative workflows (such as iterative coding, codebook development, and theme refinement) are time- and labor-intensive [8, 11, 36]. Over the last two decades, topic modeling enabled scalable text summarization and clustering, but they tend to struggle with nuance, domain shifts, and interpretability in short, heterogeneous responses [1].

Large language models (LLMs) have renewed interest in automation of qualitative content analysis by enabling flexible, instruction-driven analytic pipelines: segmenting responses into semantic units, labeling sentiment or stance, generating candidate themes, drafting codebooks, and producing human-readable summaries [23, 34]. In education, LLMs have been tested, for example, as tools to accelerate the extraction of main themes from student feedback while maintaining alignment with instructor interpretations [17]. In parallel, a broader methodological literature has emerged on LLM-as-a-judge evaluation frameworks, enabling scalable quality assurance but introducing known risks such as bias, verbosity effects, and sensitivity to prompt framing [28, 43, 48].

The application of LLMs to educational data is a growing field. Research by Katz et al. [23] has demonstrated that LLMs can be used to effectively build qualitative codebooks for student evaluations of teaching, achieving both scalability and consistency, with categories closely aligned with human-generated frameworks. Their method includes extracting units (sentences/phrases), embedding them, clustering semantically similar items, and then prompting an LLM to name clusters, write definitions, and produce a human-usable codebook. This work demonstrated significant reductions in the time and cost typically associated with manual thematic analysis. Furthermore, it addresses a limitation of purely statistical topic models (often hard-to-interpret topic-word distributions) by using LLMs to generate coherent labels and descriptions. Recent evaluation work suggests that LLM-based topic modeling can produce coherent topics but may also exhibit shortcuts (focusing on partial documents) and limited controllability via prompts, reinforcing the need for diagnostic checks and human review [16].

Parker et al. [34] showcased that with effective prompting, LLMs can achieve human-level performance in analyzing educational survey feedback. Their findings highlight the LLMs' flexibility in supporting both inductive (bottom-up) thematic analysis and deductive (top-down) classification, enabling faster, more reliable, and rigorous interpretation of student feedback than traditional methods. LLMs have also been used to generate themes for predefined prompts and compare them to instructor-derived themes, reporting substantial agreement while reducing analysis time [17]. This approach resembles assisted deductive coding, where the researcher supplies the conceptual frame, and the LLM performs standardized extraction across multiple responses.

Building on these advancements, summarised in Table 1, our study proposes and implements a comprehensive LLM-assisted feedback analytics pipeline, which includes decomposition, sentiment analysis, classification, and summarization of student open-ended responses to the course evaluation survey, specifically tailored for providing course instructors and coordinators with well structured and actionable information for course improvement. The objective is to support teachers' agency by making student feedback accessible to them and then letting them decide what is relevant and whether / how they want to act on it. In other words, the objective is to empower teachers to make use of the students' qualitative feedback that has so far been mostly ignored due to technical difficulties of natural language processing.

3. Methodology: an LLM-assisted feedback analytics pipeline

Figure 1 illustrates the data flow and key steps in the proposed feedback analytics pipeline. Each step is described in detail in the following subsections.

Table 1. Summary of related work

Objective	Analytics tasks	NLP methods / tools	Reference
Enable scalable and theoretically grounded analysis of students' open-ended feedback	Topic detection and thematic analysis	Topic modelling (LDA)	Hujala et al., 2020 [21], Sun et al., 2023 [42]
Enhance teaching practices through sentiment analysis of students' course feedback	Classification of students' comments based on the expressed sentiment	Fine-tuned language models	Pena-Torres [35]
Identify the best model for aspect-based sentiment analysis of students' comments	Aspect extraction and sentiment classification	Multi-task learning with three pre-trained language models and selective paraphrasing as a data augmentation technique	Gul et al., 2025 [18]
Enable teachers to collect, view and analyze student feedback in a more engaging and timely manner	Extraction of key themes and summarisation of student-chatbot conversations	LLM-based (GPT-4) conversational agents for student feedback collection, analysis, and presentation (via dashboard)	Marram et al., 2025 [30]
Examine and demonstrate how LLMs, employed in a variety of NLP tasks, can be used to derive insights from course feedback surveys	A variety of LLM-based NLP tasks, including classification, information extraction, thematic and sentiment analysis	LLMs (GPT-4 and GPT-3.5)	Parker et al., 2024 [34]
Examine how efficient and aligned with instructors LLMs are in identifying main themes in students' feedback	Thematic analysis and percentage agreement to determine alignment between instructors and ChatGPT	LLM-based chatbot (ChatGPT, model version not specified)	Fuller et al., 2024 [17]
Demonstrate how a combination of NLP techniques and LLMs can be used to generate a codebook for SET analysis	An NLP-based pipeline for developing a codebook for qualitative analysis of student feedback	A sequence of NLP methods: information extracting and embedding, followed by clustering of embeddings, and summarization of the clusters	Katz et al., 2024 [23]

Note: All presented studies except [30] focused on the students' open-ended responses to a course evaluation survey; the study reported in [30] differs in that its focus was on student-chatbot conversations through which course-related feedback was elicited.

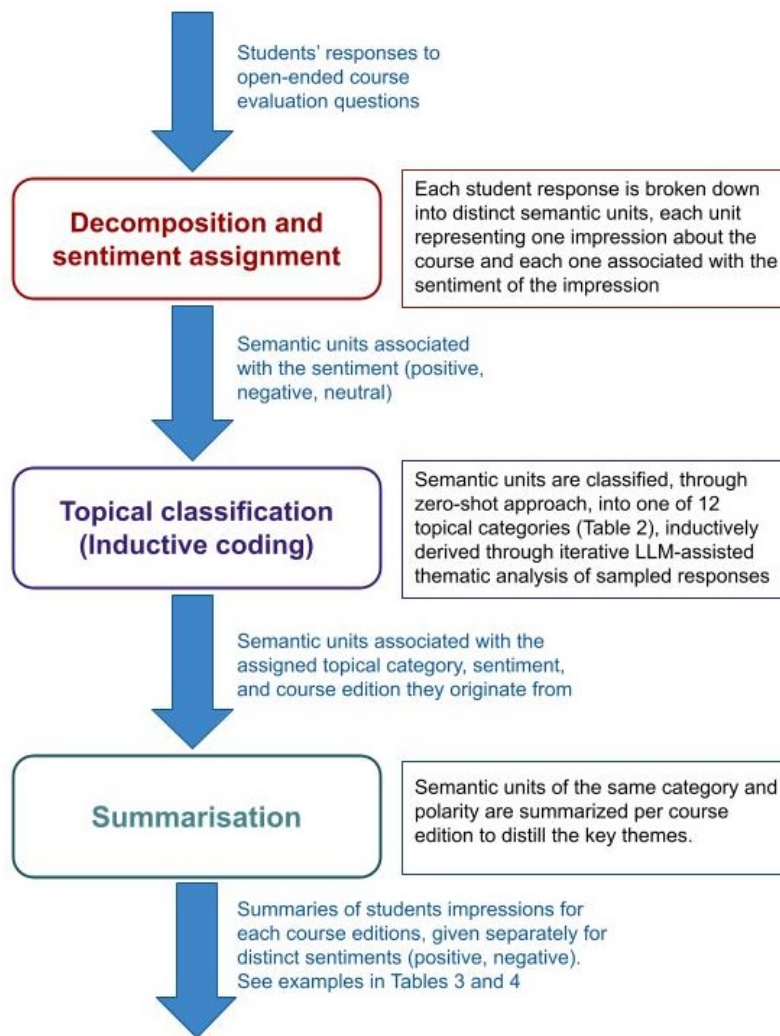


Fig. 1. Flow chart depicting the key stages of the LLM-assisted feedback analytics pipeline

3.1. Decomposition and sentiment analysis of students' comments

End-of-course surveys typically include open-ended questions that explicitly ask students what they liked and found positive about the course itself, the course instructors, or the technology used in the course. Such questions are often complemented by questions asking students about what could be improved in the course itself, in how it is taught, and in the technology supporting the course delivery. Alternatively, students may be presented with general open-ended questions asking them to comment on what they liked or disliked about the course, the instructors, or the supporting technology. This suggests that questions may have well-defined polarity (sentiment) - being either about positive or negative aspects of the course as perceived by the students - or be more broadly framed, aiming to collect students' both positive and negative sentiments about various aspects of the course. This further implies that, among students' responses, there will be responses expressing only positive sentiments, those expressing what students perceived as weaker aspects of the course, and responses expected to contain a mixture of both positive and negative comments. In addition, in all the responses, it is realistic to expect that they would often contain not an opinion on a single aspect of the course, but rather opinions about multiple distinct aspects of the course.

The above described characteristics of students' responses indicate the need to decompose each response into distinct semantic units, each one corresponding to one impression / opinion about the course. Semantic units can be of arbitrary length: a semantic unit can correspond to a sentence or can be a part of a sentence (in case of a longer sentence where a student expresses multiple opinions about a course) or consist of a few sentences (in rare cases when a student elaborates on an expressed opinion). In addition, for general questions (i.e., those without explicitly assigned polarity), the sentiment (positive, negative, neutral) of each semantic unit also needs to be determined. Accordingly, the first stage of our analytics pipeline checks whether a response contains multiple impressions (or expressions of opinion) and, if so, splits it into multiple semantic units, each one containing a single impression. In case of a response to a question with clearly defined polarity (i.e., asking students about positive or negative impressions), the extracted semantic units are assigned the same polarity as the question (e.g., positive sentiment, if the question asked about positive aspects of the course). In case of a general question, that is, a question expecting from students to share both their positive and negative impressions, each semantic unit extracted from a student response also needs to be annotated with the sentiment it expresses. To properly cover both cases, we have designed two alternative response decomposition processes:

- For responses to questions with clearly defined polarity, we prompt an LLM to identify all distinct course aspects the student was positive/negative about in their response and return them as a list of distinct impressions about the course. The LLM is also requested to avoid redundancy in the results, which may occur if a student comments on the same or very similar course aspects multiple times, by summarizing the redundant comments before including them in the result.
- For general questions without explicitly assigned polarity, in addition to the above-stated instructions, we also ask the LLM to associate each extracted semantic unit with the expressed sentiment (positive, negative, or neutral).

System prompts used in both versions of the decomposition process are provided in Figure A1 and Figure A2 of the Appendix.

3.2. Topical classification

The objective of this step is to associate each semantic unit (resulting from the previous analytics stage) with a topical category that best characterizes the student's impression, as expressed in that unit². To this end, we have defined a classification (coding) scheme as a set of topical categories with associated descriptions (Table 2), following the procedure described in the next paragraph. This scheme was used to prompt the LLM to do the classification task. Specifically, the LLM is prompted to associate the given student comment (i.e., semantic unit extracted from the student response) into one of the 11 topical categories, with each category being described in the prompt (as in Table 2). Aiming to prevent the LLM from assigning a random category when it cannot determine the correct one for a particular comment, the LLM is instructed to assign the "Other" category to a comment that does not belong to any of the given categories. The system prompt used to instruct the LLM to classify individual semantic units is given in Figure A3 of the Appendix.

The categories within the classification scheme were identified through an iterative exploratory process where the LLM was first prompted to identify the main themes within a set of semantic units (N=500) randomly sampled from student feedback in a university course that served as our initial application case (see Section 4.1). The identified themes were checked against the coding schemes used in related prior work [34] and served as the initial set of topical categories that were used to classify another random sample of semantic units (N=250), and both the misclassifications and the assignments to the "Other" category were used to revise the classification scheme and repeat the classification step. After three iterations of classification followed by revisions of the classification scheme (categories and/or descriptions), misclassifications were substantially reduced while assignments to the "Other" category were limited to those semantic units that truly could not be considered as useful (i.e., interpretable) pieces of feedback (e.g., "Time management skills", "Class preparation"). In deciding upon the classification scheme to adopt, we also consider the optimal number of categories, since a larger number of categories (i.e., more fine-grained classification) allows for more precise and nuanced organisation and analysis of student comments, but, at the same time, may negatively affect precision / correctness of the classification output (more categories tend to make the classification task more error prone). The chosen classification scheme is the one that provided a good balance between the two opposing objectives. We reflect on this further in the Discussion section. Even though this classification scheme is derived from students' comments in one course, it is rather general and may be applied, at least as a first version, to any university course for further refinement.

3.3. Summarization

The objective of this step is to create, for a particular course edition, a summary of the student comments for each topical category (Table 2) so that the course coordinator and instructors can be provided with a gist of students' impressions about distinct aspects of

² Note that, in line with the computer science terminology, we use the term "topical classification" to refer to this step of the analytics pipeline; in qualitative analysis, this task would be referred to as "inductive coding" as it includes the development of a coding scheme and the use of that scheme to code (annotate) textual content

Table 2. The coding scheme for topical classification of students' comments

Category	Description
Course content	This category refers to subject-matter topics taught in the course, including domain-specific skills and competences.
Course organization and logistics	This refers to all aspects of the course organisation, including scheduling of lectures and labs, timeliness of course announcements, the availability and accessibility of the learning tools and the like.
Course value and relevance	This refers to the perceived course value and relevance, both from the academic and practical perspective.
Course materials and resources	This category refers to all kinds of materials and resources that were provided to the students, recommended to them, or requested from them to use.
Course assignments and grading	This category refers to all kinds of assignments in the course, both formative (not graded, for practice only) and summative ones (graded), as well as the grading system applied in the course.
Instructor's teaching approach	This category refers to various aspects of the instructor's teaching approach, including different aspects of the instructional design and interaction with the students during the class.
Instructor's knowledge and expertise	This refers to the instructor's competence, as perceived by the student, for the topics included in the course syllabus.
Instructor's clarity and explanations	This category refers to the instructor's clarity of expression and quality (comprehensibility) of explanations; in other words, the quality of communication of the subject matter topics to the students.
Instructor's feedback and support	This is about the frequency, quality, and usefulness of the feedback on course assignments and the overall support.
Instructor's demeanor and communication with students	This category is about the politeness of communication with students, friendliness, and availability for communication.
Overall impression	This refers to the student's overall impression about the course and/or the instructor.

that course delivery. To that end, an LLM is given a set of student comments (semantic units) from one course edition and related to one topical category and prompted to identify the main themes in the students' comments, and for each theme, to name it and to describe it succinctly. For each course edition, the LLM is tasked with the same prompt 11 times, each time with a set of students' comments categorised into one of the 11 topical categories (those from Table 2). The system prompt used for this task is given in Figure A4 of the Appendix.

3.4. Evaluation of the pipeline

The presented feedback analytics pipeline uses an LLM for four main tasks: splitting student responses into logically and topically cohesive units (semantic units), sentiment analysis, topical classification, and summarization. The potential of LLMs for semantic text segmentation, sentiment analysis, and summarization tasks have already been explored and evaluated in prior research (e.g., [4, 22, 29, 37, 39, 44]). Therefore, we focus the evaluation primarily on the topical classification part of the pipeline, which, due to the fairly large number and closely related semantics of topical categories, may be error-prone.

To validate topical classification of students' comments (semantic units) in both scalable and trustworthy manner, we have adopted a semi-automated two-step approach based on the LLM-as-a-Judge concept [48]. LLM-as-a-Judge is about using a highly capable LLM to automatically evaluate the quality of outputs generated by another, often less capable, LLM. It offers a scalable, cost-effective alternative to human evaluation for various LLM-based tasks, including qualitative coding and text classification. The LLM-Judge is provided with: (i) the original task (in our case, topical classification of a student comment), (ii) the LLM output that is to be evaluated (the category assigned to the student comment), and (iii) the criteria to be used for the evaluation.

The first step of our evaluation approach includes both human evaluation of a randomly sampled subset of students' comments and evaluation of the same subset by an LLM prompted to serve as a Judge. The objective of this step is to ensure that the LLM-Judge can reliably validate the topical classification of student comments on a smaller sample before applying it to validate a larger sample of comments. To this end, two human experts are asked to independently evaluate the correctness of the topical classification done in stage 2 of the pipeline. Specifically, for each student comment from the randomly chosen subset, the human evaluators are instructed to either confirm that the topical category assigned by the LLM classifier is correct or to mark it as incorrect and note down the category they deem correct (out of the 11 categories in Table 2 plus the 'Other' category). After human experts complete this task and resolve any inconsistencies in their individual evaluations, the resulting joint evaluation is considered the "ground truth" to be used for assessing the correctness of the evaluation done by the LLM-judge.

The LLM-judge is given the same instructions as human evaluators, namely to verify if a student comment is associated with the correct topical category. In addition, the LLM-judge is prompted to provide the rationale for its decision. The motivation for this additional request is twofold: i) it has been shown that LLMs tend to produce better results if they are prompted to "think" before generating the output [12]; ii) to be able to trace the LLM's "thinking process", especially in erroneous cases, so that the instructions (i.e., the prompt) can be refined. The complete system prompt for the LLM-judge is given in Figure A5 of the Appendix. To determine the level of alignment between the LLM-judge and

the ground truth produced by the human experts, we use measurements that are typically used for assessing the agreement of two human raters, namely the percentage agreement and Cohen's Kappa [46].

After the LLM-judge is assessed as reliable for verifying the classification task, it is used in the second step of the evaluation process, to evaluate either the entire set of student comments or a sufficiently large (random) sample that can ensure trustworthy classification of the entire set. While evaluating the entire set of comments is preferred, its feasibility depends on the funds and infrastructure available for feedback analytics; in cases of limited resources, a randomly sampled subset may be a viable alternative (as, e.g., in [32]).

In addition to examining the correctness of the topic classification, we also examine the internal consistency of the LLM classifier. To that end, we run the classification task five times on the same random sample of student comments (semantic units) and compute the level of agreement in the classification results (i.e., percentage agreements) between each pair of runs.

4. Application case: A Financial Management course

4.1. The application context and data source

The analysis utilizes student feedback from a standalone graduate-level university course in financial management, spanning 43 semesters from 2013 to 2025. The course was run in multiple distinct course sections per semester, four semesters per year: Fall, Spring, and two Summer terms. The raw data used in this study consists of anonymised open-ended student responses from course evaluation surveys. The open-ended questions in these surveys changed a couple of times during the examined time period, but their gist remained unchanged: asking students to comment on what they liked about the course and the instructors, what they thought could be improved in the course (e.g., course content, the way it is taught, the technology used), and what was their overall impression about the course. Some of the questions were positively intoned (e.g., "What were the most valuable aspects of the course?"), others were explicitly asking for suggestions for improvements (e.g., "Please comment on aspects of this course that might be improved"), and there were also questions neutral in tone, asking students to share any thoughts they have about the course (e.g., "Feel free to provide any additional comments you believe would be useful about the course. The following topics offer suggestions, but please do not feel limited to them. What are the strongest aspects of the course overall? What can be improved? What additions/changes would you suggest?").

The original data consisted of 5892 student open-ended responses. However, some of the responses were overly brief and not bearing meaningful information. For example, responses such as "no", "nope", or "no comment" to a question asking students if they had any additional comments about the course to share. Furthermore, there were also questions that were not course-focused but asking students about their overall campus experience, their interest in other courses, or asking them about their motivation for enrolling in the given study program. After removing overly short responses and responses not directly related to the course experience, we ended up with 4679 student responses, which served as the input to our feedback analytics pipeline. From the perspective of the

above-mentioned three categories of open-ended questions, 1977 responses originated from positively intoned questions, 154 responses from negatively intoned ones (i.e., those asking for suggestions for improvement), and 2548 were given in response to the general ones (neutral in tone). The use of this data for the study presented in the paper is approved by the institution's internal review board, under the approved protocol number 8467X.

4.2. Feedback processing results

All three stages of the feedback analytics pipeline (outlined in Figure 1) were executed with the GPT-4o-mini, as a representative of the state-of-the-practice LLMs with a very reasonable price-performance trade-off. Since student survey responses were completely detached from any other student data (i.e., no student data apart from their responses to the course evaluation survey was available in the dataset) and the response data was fully anonymised, there were no privacy risks of using externally hosted LLM.

The first stage of our feedback analytics pipeline produced 9,674 student comments (semantic units) with positive sentiment, 6,044 comments with negative sentiment, and 794 neutral comments. The distribution of positive and negative comments, which were the subject of further analysis, is presented in Figure 2.

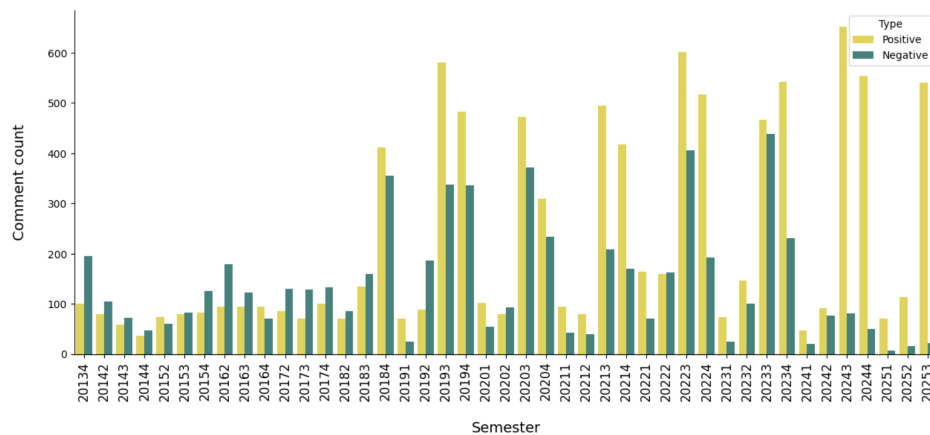


Fig. 2. Distribution of positively and negatively intoned student comments (semantic units) over the examined time period. The values on the X-axis refer to distinct, chronologically ordered course editions, consisting of the year and semester identifier (e.g., “20253” refers to the 3rd semester in year 2025).

As Figure 2 indicates, the amount of feedback collected from students differed over time, with the numbers significantly increasing for the first time in the last semester of 2018 and then oscillating over the remaining years. It is also evident that positive comments tend to outnumber negative ones, in some semesters being even many times higher than negative ones. This may be partly explained by the dynamics of student enrollment throughout an academic year. In particular, semesters labeled as 1, 2, 3, and 4 correspond

to Summer 1, Summer 2, Fall, and Spring, respectively, with the summer semesters having fewer students than the Fall and Spring ones, when the majority of students enroll in courses. After 2017, there has been a significant growth in student enrollments due to a complete overhaul of the management programs, where the course that is the focus of our study has become a core course in different study programs, including marketing, business analytics, innovation, project management, enterprise risk management, and supply chain management. Additionally, the course itself has undergone redevelopment during this transition period, which may also explain the increase in the students' positive comments and their dominance over the negative ones. The course is continually updated, as there is always room for improvement, especially in course organization, resources, and assessments, as will be confirmed by our subsequent analysis.

The next stage in the feedback processing pipeline was topical classification of positive and negative comments. The resulting distribution of positive and negative comments across the topical categories is shown in Figure 3. More precisely, the figure shows the proportions of positive and negative (i.e., improvement-oriented) comments assigned to each topical category relative to the total number of student comments. This overall distribution suggests that students had far more positive than negative overall impressions about the course during the examined time period (2013-2025). They were also predominantly positive about the course instructors. Likewise, the course content and its value and relevance were the subject of primarily positive student remarks. On the other hand, the course assignments and grading, course materials and resources, as well as the course organization and logistics received many more comments on potential improvements than praises.

While Figure 3 provides a structured insight into the students' opinion about different aspects of the course, it is limited by being based on comment counts during the overall examined time period. For course coordinators and instructors, insights into the trends over a specific time period (e.g., before and after certain instructional intervention) could be more informative and actionable. With student comments associated with the course edition they originate from and annotated with topical categories, such temporal views can be easily created. For example, Figure 4 presents one potential approach to presenting trends in students' positive opinions throughout 2024 and 2025 semesters (a similar one can be seamlessly applied to negative opinions, that is, suggested improvements). The figure shows, for each semester, how positive comments were distributed across the 11 topical categories. For example, one may notice that the second semester of 2024 ("20242") distinguishes itself by a notably higher proportion of positive overall impressions than the other semesters in 2024 and 2025, and that this may be primarily attributed to the positive impressions of course instructors.

Finally, the summarization stage produced for each course edition and each topical category a set of main themes, each one accompanied by a brief description. As an illustration of the summarization results, Table 3 presents a feedback summary for the Instructor's teaching approach category in the 2nd semester of the 2022 course edition, while Table 4 provides the same kind of summary for a course-focused category, namely Course assignments and grading, in the same course edition. The output of this phase may be particularly relevant to course instructors as it provides deep insight into distinct aspects of the course and may offer information for course and/or instructional revision or redesign.

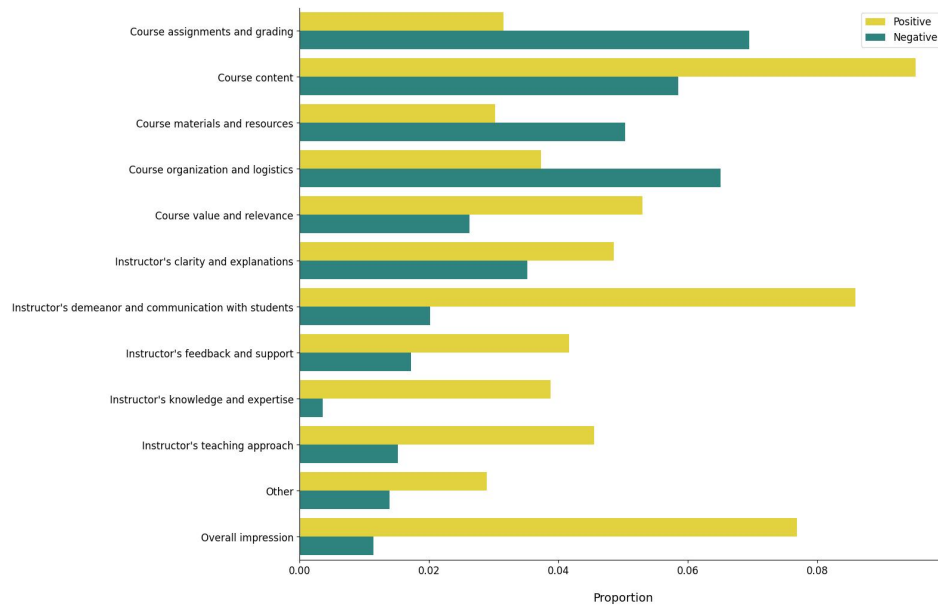


Fig. 3. Distribution of positive and negative student comments (semantic units) across the 11 topical categories plus “Other”. The values on the X-axis are the proportions of distinct categories of positive / negative comments in the total number of student comments.

Considering the complexity of a formal validation of the summarisation task, we opted for the approach adopted in similar prior studies (e.g., [34]), namely an informal evaluation by a course instructor. In particular, one of the paper co-authors, who has been teaching this course for several years, reviewed a random sample of summaries and confirmed that they accurately reflect the content of the student comments and offer actionable information relevant to the course instructors.

4.3. Evaluation of topical classification

To evaluate the topical classification of student comments (semantic units), we followed the procedure detailed in Section 3.4. For the first step of the evaluation procedure, we have randomly selected 50 positive and 50 negative (i.e., improvement-focused) student comments from each of the last five years (2021-2025). This resulted in 250 positive and 245 negative comments, for a total of 495 comments. The lower number of negative comments is due to the fact that the number of negative comments in the first three semesters of 2025 was limited to 45 (the data was not available for the 2025 Fall semester). Both positive and negative comments were evaluated by two of the paper authors. Both of them are university professors with over twenty years of experience in teaching and course management. Due to these roles, they are highly familiar with and experienced in reading and analysing student feedback. One of them has been teaching for several years the course that the comments originate from. Considering all this, the two human raters are well qualified to understand and classify students' course evaluation comments. The



Fig. 4. Distribution of students' positive comments across topical categories over semesters of 2024 and 2025 academic years. Note that all the plots have the same x-axis and to avoid the clutter, we show axis labels only for the plots in the bottom row. Likewise, labels on the y-axis are given only for the left most plots and the same y-axis applies to all the plots.

Table 3. An example of feedback summary for the Instructor's teaching approach category in one course edition

Topical category: Instructor's teaching approach	
Summary of positive comments	Instructor Effectiveness: Students praised the instructors for their thoroughness, preparation, and ability to present information clearly, contributing to a positive learning experience. Engagement in Live Sessions: Many students expressed appreciation for the live instruction and interactive homework sessions, highlighting their effectiveness in enhancing understanding of the material. Flexibility in Teaching Methods: Facilitators were commended for offering various methods of problem-solving, allowing students to choose the approach that suited them best. Course Structure: The organization and structure of the course were noted as key factors that contributed to its success and the students' learning outcomes.
Summary of negative comments	Facilitator Effectiveness: Students felt that the facilitators did not effectively engage with the material or provide adequate explanations. Need for Practical Examples: There is a desire for facilitators to incorporate real-life examples from their experiences to enhance understanding of the topics. Learning Environment: The live class environment was perceived as lacking in interactivity and depth, making it difficult for students to learn effectively. Material Presentation: Comments indicated that the material was often presented in a way that felt repetitive and not sufficiently taught, leading to confusion.

Table 4. An example of feedback summary for the Course assignments and grading category in one course edition

Topical category: Course assignments and grading	
Summary of positive comments	Homework Effectiveness: Students found the homework sessions and assignments to be beneficial for understanding course material and preparing for exams. Feedback Mechanism: The embedded homework modules with instant feedback were appreciated for enhancing learning and providing clarity on problem-solving approaches. Grading Fairness: Students felt that the grading was fair, especially given the complexity of the assignments and the tight timelines. Support Sessions: Homework help sessions and exam preparation classes were highlighted as valuable resources for reinforcing knowledge and identifying gaps. Relevance of Assignments: Assignments were seen as closely related to the course content, aiding in the overall learning experience.
Summary of negative comments	Grading Discrepancies: Students reported inconsistencies and errors in grading, leading to frustration and confusion regarding their scores. Homework Confusion: Many students found the homework assignments on Connect to be confusing and counter-intuitive, particularly with the use of Microsoft Excel. Discussion Board Issues: Feedback indicated that the discussion board grading was lacking, with no comments made and a perception that it was a waste of time. Workload Concerns: Students expressed that the workload was excessive, with too many assignments and readings that did not align well. Preference for Human Grading: There was a strong preference for assignments to be graded by a person rather than a computer, as students felt this would reduce errors and improve fairness.

same sample of positive and negative comments was also evaluated by GPT-5-mini in the LLM-Judge role (for instructions given to the model, see Section 3.4).

Human experts first evaluated the LLM classifier independently of one another, then jointly reviewed disagreements and resolved them, reaching complete agreement. Thus obtained human evaluation served as the “ground truth” for computing the evaluation measures for both the LLM classifier and LLM judge.

The number and percentage of correct topical classifications, done by the LLM classifier, as evaluated by both human experts and the LLM-judge, are given in Table 5. As the table suggests, the LLM classifier (GPT-4o-mini) tends to be more successful in classifying positive than negative comments. In addition, the LLM-judge (GPT-5-mini) seems to be quite well aligned with human experts when judging the correctness of comment classification. To further explore the alignment between the human experts and the LLM-judge, we have computed and reported the matching ratio and Cohen's Kappa in Table 6. While the matching ratios are high for both positive and negative comments, Cohen's Kappa values indicate substantial agreement for positive comments and moderate agreement for negative comments [25].

Table 5. Evaluation of the topical classification by human experts and LLM-judge, given as the number and the proportion of correct classifications

	Positive comments	Negative comments
LLM Classifier - Human	221 (0.884)	212 (0.865)
LLM Classifier - LLM Judge	223 (0.892)	211 (0.861)

Table 6. Measures of the alignment between the human experts and the LLM-judge on the evaluation of the comments classification task

	Matching ratio		Cohen's Kappa	
	Positive comments	Negative comments	Positive comments	Negative comments
LLM Judge - Human agreement	234 (0.936)	220 (0.898)	0.678	0.568

Considering this imperfect agreement, we have examined in detail the cases of misalignment between the human raters and the LLM-judge. Table 7 shows the category pairs that occurred in the misaligned cases more than once. The *Instructor's teaching approach* and *Course content* categories were the ones with the highest presence in the unaligned evaluations (23 out of 41 cases in total); likewise, *Course organization and logistics* and *Instructor's demeanor and communication with students*, occurred a couple of times. For the presented pairs, as well as those that occurred just once and are not shown here, it is clear that they are closely related in meaning, with very subtle differences, and thus

occasional misalignment is expected, especially considering that the students' comments (semantic units), when extracted from the overall student response, can be vague.

Finally, to examine the internal consistency of the LLM classifier, we reran the classification task on the same sample of 495 student comments and computed the average percentage of matches across all pairs of runs. The average matching is 0.988, with 11 to 13 mismatches out of 495 classifications, per run. This indicates high internal consistency of the LLM classifier.

Table 7. Frequent kinds of misalignment between human raters and LLM-judge

Category 1	Category 2	Frequency
Instructor's teaching approach	Course organization and logistics	9
	Instructor's demeanor and communication with students	2
	Instructor's clarity and explanations	2
Course content	Course value and relevance	4
	Course organization and logistics	3
	Course materials/resources	3
Instructor's feedback and support	Instructor's demeanor and communication with students	3
Course assignments and grades	Course organization and logistics	2

5. Discussion

This paper presents a practical, LLM-based methodology for scalable analysis and information extraction from students' open-ended feedback, collected via an end-of-course evaluation survey. Through the main steps of the proposed feedback analytics pipeline (Figure 1), students' feedback, from one or a series of course editions, is transformed from raw natural language responses, to a set of main themes, each associated with one of the eleven distinct course or instructor-related aspects of a course evaluation (Table 2). This output may serve as a valuable source of information to instructors and course coordinators for course and instruction (re-)design. It may also offer course coordinators and school management insights into trends in students' opinions about specific courses and instructional aspects over longer periods, and thereby inform course-level and program-level decisions.

The presented application case (Section 4) illustrates the method's output in the context of a graduate-level university course. Specifically, Tables 3 and 4 illustrate summaries, in the form of a set of key themes, that the method generates for each of the eleven course aspects that emerged as relevant in the overall pool of student responses. While these summaries may be primarily relevant to the course instructors and facilitators who taught the corresponding course edition, a visual depiction of the change in the students' opinion over time, such as the ones shown in Figures 3 and 4, may help the course coordinator to better understand the effects of the applied course / instructional changes over time.

This work complements prior related work, especially the work by Fuller et al. [17] and Katz et al. [23]. The former examined how efficient and aligned with instructors LLMs were in identifying main themes in students' feedback, whereas the latter demonstrated how a combination of NLP techniques and LLMs could be used to generate a codebook for student feedback analysis. In the current work, we show how a codebook, once developed, can be systematically employed in a feedback analytics pipeline to identify main topical categories and use these to prepare topically-focused feedback summaries for instructors. Our work is most similar to the work of Parker et al. [34] who did a comprehensive exploration of LLMs use in a variety of NLP tasks, for deriving insights from course feedback surveys. While they have explicitly covered a wider span of potential scenarios for leveraging student feedback, our solution has the advantage of being more versatile due to the initial step of decomposition of the feedback into semantic units, which together with the assigned sentiment may serve as suitable input for a variety of feedback analytics scenarios.

The evaluation of the classification step (Section 4.3) has shown that neither the LLM-classifier nor the LLM-judge is fully aligned with the human experts. This is in line with recent research that studied LLM annotation errors in detail and offered empirical evidence for why a high level of alignment between humans and LLMs on qualitative coding tasks is unrealistic [45]. However, in the majority of misclassification cases, the errors we observed are not "consequential" in the sense that they are not expected to have major impact on the downstream tasks, namely, topical summaries to be created for teachers and/or visualizations for course/study program coordinators. For example, a typical misclassification was that the LLM assigned the *Course content* category instead of *Course value and relevance*, or vice versa (Table 7). Considering how closely related these two categories are, whether a student comment is included in the summary of one or the other category, is not expected to affect much the quality of information that is eventually provided to the teacher. Similarly, such a misclassification would not significantly affect the trends in the student opinion that a visualization such as the one in Figure 4 aims to communicate.

The observed misclassifications (i.e., errors of the LLM-classifier) and misalignments with the LLM-judge, were to be expected since, in practice, the agreement among human classifiers can be imperfect too. Human classifiers and judges may encounter ambiguities and/or unclear statements in student comments (e.g., "Very responsible", "Grading if discussions are unclear and misleading"), which may occur due to students' writing styles, their command of the language if they are non-native speakers, and so on. Such ambiguities may also be a source of uncertainties for human evaluators regarding how the comments should be best classified. In the presented application case, such situations

were not uncommon (15-20%), depending on the specific human evaluator, and were eventually resolved through a discussion among the human evaluators.

Uncertain human classification of student opinions and misalignments between human and LLM classifiers and judges are tightly coupled with the coding scheme complexity [45]. On the one hand, the more complex the scheme, the greater the likelihood that student comments will convey the message precisely. On the other hand, a complex classification scheme tends to introduce uncertainties about the right category. This may create a preference for multi-label coding, that is, the use of two or more categories to label students' comments (as done, e.g., in [34]), or for more detailed instructions for the LLM classifier to avoid typical inconsistencies and ambiguities. Further research is required to systematically investigate whether multi-label classification and/or more detailed instructions would achieve better overall results than a single-label.

6. Limitations

Like any other research method and study, the current one suffers from some limitations.

First, the approach presented in the paper was evaluated with data from only one course, although the dataset of student comments was quite large and spanned multiple course editions over twelve years. For a more comprehensive evaluation, similar datasets from other courses should be used as well.

Second, the only component of the feedback analytics pipeline thoroughly evaluated in this study is the topical classification (Section 4.3), which was conducted on a relatively small subset of student comments ($N=495$). For a fully trustworthy feedback analytics pipeline, it is necessary to also evaluate how accurately the LLM extracts semantic units from student comments, how well it performs sentiment analysis, and how well the generated summaries align with students' comments. While prior research has evaluated LLMs on tasks of text segmentation, sentiment analysis, and summarization, these evaluations were conducted in different contexts, and it would also be important to verify that LLMs can complete these tasks well in the context of students' open-ended course feedback.

Next, all the tasks in the analytics pipeline were based on one specific LLM, namely GPT4-o-mini, whereas GPT5-mini served as the LLM judge. While these models have been representatives of the state-of-the-practice LLMs at the time the study was conducted, testing the pipeline with other LLMs, especially open-weighted ones is an important next step, as some universities may prefer to use locally hosted models for student feedback analysis.

Finally, our evaluation of the LLM-judge showed either substantial or moderate agreement with human experts. Systematic analyses of LLM-as-a-judge highlight that even strong judge models may lag inter-human agreement and can display vulnerabilities such as leniency, bias, and sensitivity to prompt complexity [43]. To address these vulnerabilities, we will explore safeguards recommended for qualitative applications, including: (i) using multiple judges or model families, (ii) further calibration on a human-labeled reference set, including the documentation of prompt versions and sampling strategies [43,48].

7. Conclusion and future research directions

Summarizing students' comments and opinions from open-ended questions in end-of-the-course surveys into meaningful and actionable insights is an extremely challenging and demanding task for humans. Using LLMs to help perform this task (semi-)automatically has started to attract interest among researchers, teachers, and course managers. Yet, until now, a generic methodology for conducting this task using LLMs was missing.

This paper has analyzed existing approaches, identified emerging common points, elaborated on them, and defined a feedback analytics pipeline for using LLMs to perform this task. The main steps in the pipeline include splitting student responses into logically and topically cohesive units (semantic units), sentiment analysis, topical classification, and summarization. The major benefit of this framework is that it enables teachers and course coordinators to start with students' qualitative feedback to open-ended questions in their course evaluations, and obtain structured and actionable insights. We have evaluated the framework in practice using a rich dataset of student comments from multiple editions of a graduate course in financial markets taught at a U.S. university and obtained encouraging results.

The presented analytics pipeline and the lessons learned from its initial application, provide grounds for some interesting further research to strengthen the validity and practical utility of the pipeline. First, the imperfect alignment between LLMs and human evaluators suggests that the current LLMs are not (yet) ready for automated qualitative coding of student open-ended feedback and that some form of human-LLM collaboration on the coding tasks might be a better solution. Accordingly, an important direction for further research is to redesign some elements of the proposed methodological pipeline, to ground them in human-in-the-loop and/or human-on-the-loop principles, similar to what was recently proposed by Li et al. [26] for coding human-AI interaction in learning settings. This would ensure the preservation of interpretive validity, as an intricate feature of traditional approaches to qualitative analysis [8,36], while leveraging scale, speed, and consistency that automation brings. Second, the adopted coding scheme (Table 2) is not grounded in educational theories but inductively created. As such, the coding scheme does not bind the resulting annotations/classifications of student comments to any particular pedagogical approach. On the other hand, being pedagogy-agnostic, the coding scheme does not provide teachers with direct guidance on how to act upon the insights derived from the students' comments. Therefore, it would be worthwhile to explore a more pedagogy-informed version of the coding scheme, first in terms of LLM-based annotation/classification accuracy, and then in terms of the utility it can bring to teachers as a part of the overall analytics pipeline. The recently proposed coding instrument for labeling effective and malfunctioning teaching practices [2] may provide a solid ground for pursuing this research direction. Last but not least, we intend to evaluate the outputs of the proposed feedback analytics pipeline with the instructors and facilitators teaching the course from our first application case. The collected educators' feedback on the relevance and usefulness of the topical summaries and visual presentation of trends in students' opinions, will serve two important functions: (i) it will serve as a key input for further refining the methodological pipeline, and (ii) it will be used for the initial design of a teacher-facing dashboard, computationally grounded in the feedback analytics pipeline, which would allow teachers to choose and shape the analysis of the course feedback.

Acknowledgment. This work was supported by the Serbian Academy of Sciences and Arts grant No. F-150.

References

1. Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., Hassan, A.: Topic modeling algorithms and applications: A survey. *Information Systems* 112, 102131 (2023)
2. Astudillo, G., Hilliger, I., Baier, J.: Unveiling teaching practices in higher education: Leveraging open-ended student comments in learning analytics. In: *Proceedings of the 16th International Learning Analytics and Knowledge Conference. LAK '26*, Association for Computing Machinery, New York, NY, USA (2026)
3. Barany, A., Nasir, N., Porter, C., Zambrano, A.F., Andres, A.L., Bright, D., Shah, M., Liu, X., Gao, S., Zhang, J., Mehta, S., Choi, J., Giordano, C., Baker, R.S.: ChatGPT for Education Research: Exploring the Potential of Large Language Models for Qualitative Codebook Development. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) *Artificial Intelligence in Education*. pp. 134–149. Springer Nature Switzerland, Cham (2024)
4. Basyal, L., Sanghvi, M.: Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models (Oct 2023), <http://arxiv.org/abs/2310.10449>, arXiv:2310.10449 [cs]
5. Berk, R.A.: Survey of 12 strategies to measure teaching effectiveness. *International journal of teaching and learning in higher education* 17(1), 48–62 (2005)
6. Biggs, J., Tang, C., Kennedy, G.: *Teaching for quality learning at university 5e*. McGraw-hill education (UK) (2022)
7. Bovill, C., Cook-Sather, A., Felten, P.: Students as co-creators of teaching approaches, course design, and curricula: Implications for academic developers. *International Journal for Academic Development* 16(2), 133–145 (2011)
8. Braun, V., Clarke, V.: Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2), 77–101 (2006)
9. Brockx, B., Van Roy, K., Mortelmans, D.: The student as a commentator: Students' comments in student evaluations of teaching. *Procedia - Social and Behavioral Sciences* 69, 1122–1133 (2012)
10. Brookfield, S.D.: *Becoming a critically reflective teacher*. John Wiley & Sons (2017)
11. Charmaz, K.: *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. SAGE (2006)
12. Chiang, C.H., Lee, H.y.: A Closer Look into Using Large Language Models for Automatic Evaluation. In: Bouamor, H., Pino, J., Bali, K. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2023*. pp. 8928–8942. Association for Computational Linguistics, Singapore (Dec 2023), <https://aclanthology.org/2023.findings-emnlp.599/>
13. Cohen, P.A.: Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education* 13(4), 321–341 (1980)
14. Cook-Sather, A.: Sound, presence, and power: “student voice” in educational research and reform. *Curriculum Inquiry* 36(4), 359–390 (2006)
15. Diamond, M.R.: The usefulness of structured mid-term feedback as a catalyst for change in higher education classes. *Active Learning in Higher Education* 5(3), 217–231 (2004)
16. Doi, T., Isonuma, M., Yanaka, H.: Comprehensive evaluation of large language models for topic modeling. arXiv:2406.00697 (2024)
17. Fuller, K., Morbitzer, K.A., Zeeman, J.M., Persky, A., Savage, A., McLaughlin, J.E.: Exploring the use of ChatGPT to analyze student course evaluation comments. *BMC Medical Education* 24(1), 423 (Apr 2024), <https://link.springer.com/article/10.1186/s12909-024-05316-2>

18. Gul, S., Asif, M., Amin, F.E., Saleem, K., Imran, M.: Advancing Aspect-Based Sentiment Analysis in Course Evaluation: A Multi-Task Learning Framework With Selective Paraphrasing. *IEEE Access* 13, 7764–7779 (2025), <https://ieeexplore.ieee.org/document/10833605>
19. Heffernan, T.: Sexism, racism, prejudice, and bias: A literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assessment & Evaluation in Higher Education* (2021)
20. Hou, C., Zhu, G., Zheng, J., Zhang, L., Huang, X., Zhong, T., Li, S., Du, H., Ker, C.L.: Prompt-based and Fine-tuned GPT Models for Context-Dependent and -Independent Deductive Coding in Social Annotation. In: *Proceedings of the 14th Learning Analytics and Knowledge Conference*. pp. 518–528. LAK '24, Association for Computing Machinery, New York, NY, USA (Mar 2024), <https://doi.org/10.1145/3636555.3636910>
21. Hujala, M., Knutas, A., Hynninen, T., Arminen, H.: Improving the quality of teaching by utilising written student feedback: A streamlined process. *Computers & education* 157, 103965 (2020)
22. Häglund, E., Björklund, J.: Opinion Units: Concise and Contextualized Representations for Aspect-Based Sentiment Analysis. In: Johansson, R., Stymne, S. (eds.) *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*. pp. 230–240. University of Tartu Library, Tallinn, Estonia (Mar 2025), <https://aclanthology.org/2025.nodalida-1.24/>
23. Katz, A., Gerhardt, M., Soledad, M.: Using generative text models to create qualitative codebooks for student evaluations of teaching. *arXiv:2403.11984* (2024)
24. Kirk, M.: Student voice and agency for transformative change in matters that matter: Impactful inquiry in primary science. *The Australian Educational Researcher* 52(1), 411–430 (May 2024), <https://link.springer.com/article/10.1007/s13384-024-00721-4>
25. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics* 33(1), 159–174 (Mar 1977)
26. Li, Z., Tang, L., Xia, M., Li, X., Chen, N., Gašević, D., Fan, Y.: When llms fall short in deductive coding: Model comparisons and human–ai collaboration workflow design. In: *Proceedings of the 16th International Learning Analytics and Knowledge Conference*. LAK '26, Association for Computing Machinery, New York, NY, USA (2026)
27. Liu, X., Zambrano, A.F., Baker, R.S., Barany, A., Ocumpaugh, J., Zhang, J., Pankiewicz, M., Nasir, N., Wei, Z.: Qualitative Coding with GPT-4. *Journal of Learning Analytics* 12(1), 169–185 (Mar 2025), <https://learning-analytics.info/index.php/JLA/article/view/8575>
28. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv:2303.16634* (2023)
29. Maraj, A., Martin, M.V.: Context is the Key for LLM-Based Text Segmentation. In: *Natural Language Processing and Information Systems: 30th International Conference on Applications of Natural Language to Information Systems, NLDB 2025, Kanazawa, Japan, July 4–6, 2025, Proceedings, Part I*. pp. 142–155. Springer-Verlag, Berlin, Heidelberg (2025)
30. Maram, S.S., Zaman, U., El-Nasr, M.S.: Listening with language models: Using llms to collect and interpret classroom feedback (2025), <https://arxiv.org/abs/2508.11707>
31. Marsh, H.W.: Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In: *The scholarship of teaching and learning in higher education: An evidence-based perspective*, pp. 319–383. Springer (2007)
32. McNichols, H., Ikram, F., Lan, A.: The studychat dataset: Analyzing student dialogues with chatgpt in an artificial intelligence course. In: *Proceedings of the 16th International Learning Analytics and Knowledge Conference*. LAK '26, Association for Computing Machinery, New York, NY, USA (2026)

33. Nasser-Abu Alhija, F., Fresko, B.: Student evaluation of instruction: What can be learned from students' written comments? *Studies in Educational Evaluation* 35, 37–44 (2009)
34. Parker, M.J., Anderson, C., Stone, C., Oh, Y.: A large language model approach to educational survey feedback analysis. *International Journal of Artificial Intelligence in Education* 35(3), 444–481 (2025)
35. Peña-Torres, J.: Towards an improved of teaching practice using Sentiment Analysis in Student Evaluation. *Ingeniería y Competitividad* 26 (Jun 2024)
36. Saldaña, J.: *The Coding Manual for Qualitative Researchers*. SAGE Publications Ltd, Thousand Oaks, California (2021)
37. Satya Putra, A., Menekse, M., Butt, A.A.: WIP: Leveraging LLM for Sentiment Analysis of Student Reflection Texts from a Large Undergraduate Course. In: *Proceedings of the Twelfth ACM Conference on Learning @ Scale*. pp. 266–270. L@S '25, Association for Computing Machinery, New York, NY, USA (Jul 2025), <https://doi.org/10.1145/3698205.3733937>
38. Schreier, M.: *Qualitative Content Analysis in Practice*. SAGE (2012)
39. Shah, F.A., Sabir, A., Sharma, R., Pfahl, D.: How Effectively Do LLMs Extract Feature-Sentiment Pairs from App Reviews? In: Hess, A., Susi, A. (eds.) *Requirements Engineering: Foundation for Software Quality*. pp. 123–138. Springer Nature Switzerland, Cham (2025)
40. Spooren, P., Brockx, B., Mortelmans, D.: On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research* 83(4), 598–642 (2013)
41. Stark, P.B., Freishtat, R.: An evaluation of course evaluations. Technical report, University of California, Berkeley (2014), <https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1>, accessed 2026-01-10
42. Sun, J., Yan, L.: Using topic modeling to understand comments in student evaluations of teaching. *Discover Education* 2, 25 (2023)
43. Thakur, A.S., Choudhary, K., Ramayapally, V.S., Vaidyanathan, S., Hupkes, D.: Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In: *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*. pp. 404–430 (2025)
44. Xiao, Z., Yuan, X., Liao, Q.V., Abdelghani, R., Oudeyer, P.Y.: Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In: *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. pp. 75–78. IUI '23 Companion, Association for Computing Machinery, New York, NY, USA (Mar 2023), <https://doi.org/10.1145/3581754.3584136>
45. Xu, Z., Khatri, V., Dai, X., Liu, X., Li, S., Zhang, X., Yu, R.: Enhancing llm-based data annotation with error decomposition. In: *Proceedings of the 16th International Learning Analytics and Knowledge Conference*. LAK '26, Association for Computing Machinery, New York, NY, USA (2026)
46. Zapf, A., Castell, S., Morawietz, L., Karch, A.: Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology* 16(1), 93 (Aug 2016), <https://doi.org/10.1186/s12874-016-0200-9>
47. Zhang, M., Lindsay, E.D., Thorbensen, F.B., Poulsen, D.B., Bjerva, J.: Leveraging Large Language Models for Actionable Course Evaluation Student Feedback to Lecturers (Jul 2024), <http://arxiv.org/abs/2407.01274>, arXiv:2407.01274 [cs]
48. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. pp. 46595–46623. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (Dec 2023)
49. Zipsper, N., Kurochkin, D., Yu, K.W., Mincieli, L.A.: Exploring relationships between qualitative student evaluation comments and quantitative instructor ratings: A structural topic modeling framework. *Education Sciences* 15(8), 1011 (2025)

Appendix

This section presents the final version of the system prompts used in distinct stages of the proposed feedback analytics pipeline.

```
You are an expert in analysing student responses to open-ended questions in a course
evaluation survey. Your specialty is extracting, from students' responses, opinions that
the students expressed about the course (e.g., about the course structure, organisation,
materials, the teacher, and the like).

I will give you a student response to a survey question asking them to comment on the
positive aspects of the course or the instructor.
In their response, a student may express a positive opinion about just one aspect of the
course and/or instructor or may express opinion about several aspects.
Your task is to identify all distinct course aspects the student was positive about in
their response and return them as a list, where each opinion is one list element.
Please try to avoid redundancy in this list, so if a student comments on the same or very
similar course aspect twice or more times, summarise those comments and then add them to
the list.

Structure the output as a JSON list of string values.

Here is the student response to analyze:
{response}
```

Fig. A1. The final version of the system prompt for decomposing students' responses to questions with explicitly defined polarity

```

You are an expert in analysing student responses to open-ended questions in a course
evaluation survey. Your specialty is extracting, from a student's response, positive
and negative aspects of the course (e.g., the course content, structure, materials, the
instructor's teaching style and approach, and the like).

I will give you a student response to a survey question asking them to comment on what
they liked about a course and the course instructor, and, also, what are the things
that should be improved regarding the course and the instructor.
In their response, a student may express opinion about just one aspect of the course
and/or instructor or may express opinion about several aspects.
Your task is to identify all distinct opinions about the course or instructor that the
student expressed in their response and determine the sentiment of the opinion
(positive, neutral or negative).
Your response should be a JSON list of dictionaries, where each dictionary is of the
following form:

{{
  'opinion': <an extracted student's opinion goes here>
  'sentiment': <the sentiment of the opinion goes here>
}}

Here is the student response to analyze:
{response}

```

Fig. A2. The final version of the system prompt for decomposing students' responses to general questions, where a response may contain a mixture of positive and negative impressions

You are an expert in analysing student responses to open-ended questions in a course evaluation survey.
 Your specialty is classifying students' comments into topical categories representing students' opinions about distinct aspects of the course and its instructors.

I will give you a student's comment on the positive aspect(s) of the course or the instructor. Your task is to classify the comment into one of the eleven categories given below.
 If the comment does not belong to any of the given categories, put it into the 'Other' category.

Here is the list of categories, where each category label is followed by a description:

- * Course content: this category refers to subject-matter topics taught in the course, including domain-specific skills and competences.
- * Course organization and logistics: this refers to all aspects of the course organisation, including scheduling of lectures and labs, timeliness of course announcements, the availability and accessibility of the learning tools and the like.
- * Course value and relevance: this refers to the perceived course value and relevance, both from the academic and practical perspective.
- * Course materials and resources: this category refers to all kind of materials and resources that were provided to the students, recommended to them, or requested from them to use.
- * Course assignments and grading: this category refers to all kinds of assignments in the course, both formative (not graded, for practice only) and summative ones (graded), as well as the grading system applied in the course.
- * Instructor's teaching approach: this category refers to various aspects of the instructor's teaching approach, including different aspects of the instructional design and interaction with the students during the class.
- * Instructor's knowledge and expertise: this refers to the instructor's competence, as perceived by the student, for the topics included in the course syllabus.
- * Instructor's clarity and explanations: this category refers to the instructor's clarity of expression and quality (comprehensibility) of explanations; in other words, the quality of communication of the subject matter topics to the students.
- * Instructor's feedback and support: this is about the frequency, quality, and usefulness of the feedback on course assignments and the overall support.
- * Instructor's demeanor and communication with students: this category is about the politeness of communication with students, friendliness, and availability for communication.
- * Overall impression: this refers to the student's overall impression about the course and/or the instructor.

Return your response as a string that contains only the label of the category you assigned to the student's comment. Do not include any preamble or explanation.

Student's comment:
 {comment}

Fig. A3. The final version of the system prompt for classifying students' comments (semantic units) with positive polarity. Note: The prompt used for categorization of semantic units with negative polarity differs only in the opening sentence of the second paragraph, which reads as follows: "I will give you a student's comment expressing dissatisfaction with some aspect(s) of the course or the instructor. The dissatisfaction may be openly/directly expressed, or it may be stated in a tacit/implicit form."

```
You are an expert in analysing student responses to open-ended questions in a course
evaluation survey.
Your specialty is summarising students' comments, to make it easier for the teacher to
identify the key themes in the students' feedback.

I will give you a set of students' comments (each comment in a separate line) about a
particular aspect of the course or the course instructors and will ask you to create a
concise summary of those comments, by identifying the key themes in the comments.

Present the identified main themes as a dictionary where each dictionary key is a concise
expression of an identified main theme, while the value is a sentence-long description of
that theme.
Return the results in the JSON format, as a list of key-value pairs.

Student comments:
{comments}
```

Fig. A4. The final version of the summarization prompt

You are an expert evaluator for text classification tasks. Your goal is to act as a judge and assess the accuracy of a student opinion classification.

Here is the task: A student's comment about a course was classified into one of the following categories:

- * Course content: this category refers to subject-matter topics taught in the course, including domain-specific skills and competences.
- * Course organization and logistics: this refers to all aspects of the course organisation, including scheduling of lectures and labs, timeliness of course announcements, the availability and accessibility of the learning tools and the like.
- * Course value and relevance: this refers to the perceived course value and relevance, both from the academic and practical perspective.
- * Course materials and resources: this category refers to all kinds of materials and resources that were provided to the students, recommended to them, or requested from them to use.
- * Course assignments and grading: this category refers to all kinds of assignments in the course, both formative (not graded, for practice only) and summative ones (graded), as well as the grading system applied in the course.
- * Instructor's teaching approach: this category refers to various aspects of the instructor's teaching approach, including different aspects of the instructional design and interaction with the students during the class.
- * Instructor's knowledge and expertise: this refers to the instructor's competence, as perceived by the student, for the topics included in the course syllabus.
- * Instructor's clarity and explanations: this category refers to the instructor's clarity of expression and quality (comprehensibility) of explanations; in other words, the quality of communication of the subject matter topics to the students.
- * Instructor's feedback and support: this is about the frequency, quality, and usefulness of the feedback on course assignments and the overall support.
- * Instructor's demeanor and communication with students: this category is about the politeness of communication with students, friendliness, and availability for communication.
- * Overall impression: this refers to the student's overall impression about the course and/or the instructor.

Finally, if a comment could not be classified into any of the above categories, it would be assigned to the "Other" category.

Both the student's comment and the category it was assigned are given below. Your task is to determine if the assigned category is correct.

The assigned category is considered correct if the following two criteria are satisfied:

- 1) Among the possible categories (given above), it is the best fit for the student's comment
- 2) The comment primarily focuses on the aspect identified by the category.

Your response should be a dictionary of the following form:

```
{
  'correctness': [Yes/No]
  'reasoning': [Brief explanation of your evaluation]
}
```

Return this dictionary as a JSON object. Do not include any preamble or any additional explanation.

Fig. A5. The final version of the system prompt used to instruct the LLM-Judge to verify topical classification of student comments

Jelena Jovanovic is a Professor at the Department of Software Engineering, University of Belgrade, Serbia. She is also an Adjunct Professor in the Centre for the Science of Learning & Technology (SLATE) at University of Bergen, Norway, and an Adjunct Professor in the Centre for Learning Analytics at Monash (CoLAM), Monash University, Australia. Her current research focus is on the use of computational approaches, including statistical and machine learning methods and techniques, network analysis, and generative AI, towards better understanding and supporting learning, primarily self-regulated learning in higher education settings. Currently she serves as Advisor to the Editor-in-Chief of IEEE Transactions on Learning Technologies and Editorial Board member of Journal of Learning Analytics, Journal of Educational Data Mining, The Internet and Higher Education, and Computers & Education: Artificial Intelligence.

Irena Vodenska is a Professor of Finance, Director of Finance programs, and chair of the Administrative Sciences Department at Metropolitan College, Boston University. Since 2004, she has been a Chartered Financial Analyst (CFA) charter holder. Her research focuses on network theory and complexity science, sustainability in finance and macroeconomics, effects of media on financial markets, technology-aided learning, and AI-powered enhancement in teaching. She is involved in Financial Literacy initiatives and democratizing finance education for the global community. As a Principal Investigator (PI) for Boston University, she has won interdisciplinary research grants from the European Commission, the US Army Research Office, and the National Science Foundation, USA.

Vladan Devedzic is a Professor of Computer Science and Software Engineering at the University of Belgrade, Faculty of Organizational Sciences, Belgrade, Serbia. Since 2021, he has been a Corresponding Member of the Serbian Academy of Sciences and Arts (SASA). His long-term professional objective is to bring close together ideas from the broad fields of Artificial Intelligence / Intelligent Systems and Software Engineering. His current professional and research interests include artificial intelligence, programming education, software engineering, intelligent software systems and technology-enhanced learning (TEL). He is the founder and the Chair of the GOOD OLD AI research network.

Received: January 20, 2026; Accepted: June 3, 2026.