

Student Perceptions of Essay Grading with Artificial Intelligence in Higher Education

Umair Ali Khan

`umairali.khan@haaga-helia.fi`

Haaga-Helia University of Applied Sciences

Ari Alamäki

Haaga-Helia University of Applied Sciences

Stephan Schlögl

Management Center Innsbruck

Research Article

Keywords: AI-based assessment, automated essay grading, user experience, student perceptions, educational AI, feedback systems

Posted Date: September 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10382929/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Assessment plays a key role in education, as it is central to both students' learning effectiveness and teachers' workload. The advantages of AI technologies in assessment have therefore attracted significant interest among researchers and practitioners across educational contexts. This paper examines students' initial experiences with AI-based assessment and essay grading in authentic classroom settings, where they gained direct experience with an AI-based essay-grading tool. An exploratory computational text analysis was conducted on 109 open-ended student responses collected after an authentic classroom experiment in Austria and Finland. The analysis combined LLM-assisted response condensation, K-means clustering, and sentiment classification. Seven thematic clusters were identified, capturing students' perceptions of the AI-based assessment and grading system: Mixed Opinions on AI Grading Effectiveness; Overall Experience with the AI Tool; User Experience and Convenience; User Trust and Skepticism; Need for Human Touch in Feedback; Feedback Quality and Improvement; and Evaluation Accuracy Concerns. Students appreciated the immediacy of feedback and the convenience of AI-based assessment but expressed concerns about evaluation accuracy, trustworthiness, and the lack of personalized, human-like feedback. The findings highlight the need for human–AI hybridity in assessment practices and emphasize the importance of orchestrating the roles of AI, teachers, assessment formats, and learning processes. Educators should balance automated feedback with human oversight when designing implementation strategies for AI-based assessment in both formative and summative contexts. The identified clusters provide a novel framework for the design and evaluation of AI-enhanced assessment practices.

1. Introduction

AI-enhanced automatic assessments are deemed to play an important role in education (Alfaleh, 2026; Deepshikha, 2026; Minn, 2022). Thus, automated essay grading systems have increasingly been gaining the interest of researchers and education providers (Gao et al., 2024; Ramesh & Sanampudi, 2022). Yet, despite numerous investigations focusing on automated essay grading, there remains a scarcity of research investigating students' user experiences with AI-based essay grading systems in higher education. In the past, automated assessment and grading systems have been simple, with only a few of them facilitating the development of complex thinking and reasoning through feedback (Gao et al., 2024).

Nowadays, AI-based grading systems are able to evaluate ill-structured and complex essays that demonstrate higher-order thinking (Alamäki, et al., 2024). Thus, they can automate traditional work duties usually performed by teachers. In doing so, the features of AI-based software systems, such as automatic feedback and responses and machine-generated content, are considered an integral part of the students' usage experience. However, as such feedback will be provided by machines instead of humans, there is a need to investigate students' perceptions of these interactions. What is more, educators and teachers lack effective design and implementation strategies for AI tools, and the

effectiveness of developing cognitive processes and skills varies significantly across different educational AI solutions (Luo et al., 2025).

Thus, for this investigation, we created a research setting, which allowed us to study how the features of advanced AI-based assessment and essay grading systems affect students' user experiences while using a system in an actual learning context. The goal was to increase our understanding of current higher education students' experiences related to educational AI. Today's students have already started to adopt AI-based systems in their studies. Understanding their preferences related to those new technologies helps teachers, trainers, and curriculum designers to utilize more AI-enhanced learning and teaching methods and design effective implementation strategies. Also, for those developing new educational AI solutions, understanding students' first impressions and user experiences regarding educational technologies is crucial. In fact, it has been shown that affective evaluation plays a key role in a user's experience (Dirin et al., 2018; Knutson & Greer, 2008). In other words, it is important to understand users' experiences as this knowledge assists in developing and implementing technological solutions that diffuse more quickly among users (e.g., Hallikainen et al., 2019; Jahanmir & Cavadas, 2018). To this end, prior research (e.g., Alamäki, Dirin & Suomala, 2021) has also pointed out that understanding students' preferences helps identify potential opportunities and obstacles regarding the future adoption of educational technologies. Consequently, to increase our understanding of students' experiences with an AI-based assessment and essay-grading system, we conducted an exploratory computational text analysis within an authentic classroom experiment. The goal was to identify the distinct user-experience issues that emerged immediately after students used the system and to examine how these issues could be grouped into semantically related clusters. The study was conducted in Austria and Finland, where students engaged with a commercial AI-based assessment and essay grading system as part of an authentic learning assignment. The respective investigation was guided by the following research question:

What are the distinct user experience issues that are raised by students immediately after using an AI-based essay grading system?

Reporting this investigation, the remainder of this article is organized as follows: Section 2 starts with a review of the relevant literature. Subsequently, Section 3 describes the methodology and the data used in the study. Then, Section 4 outlines our findings before Section 5 discusses the study's contributions and its theoretical implications. Finally, Section 6 outlines the study's limitations and points to potential future research directions.

2. Related Work

2.1 AI-based Assessment in Education

Automatic computer-based assessment systems evaluate students' written responses (Ramesh & Sanampudi, 2022). Traditionally, computer-based assessment has been used to evaluate and grade

multiple-choice questions and various quizzes, or to calculate the number of characters or keywords in written responses. The adoption of AI technologies in automatic computer-based assessment has enabled more sophisticated evaluation mechanisms that are capable of handling complex and large text-based essays and written responses. Although teachers are still responsible for the final grading of courses in higher education of many countries, AI-based assessment can provide instant learning feedback to students on their assignments and support teachers in delivering informal assessment and feedback, which is crucial for effective learning. Thus, AI-based assessment and grading are gradually being integrated into teaching practices, automating certain aspects of assessment in education (Alamäki et al., 2024; Deepshikha, 2026). They have also been increasingly attracting the interest of researchers (Alfaleh, 2026; Gao et al., 2024; Ramesh & Sanampudi, 2022), as automatic and AI-based assessment is seen as a key enabler for the automation and digitization of educational practices. Therefore, it is important to study students' initial experiences when engaging with automatic and AI-based assessment and grading in authentic classroom settings, where AI tools are used in real-life learning and teaching situations.

2.2 Forming First Impressions in Human-AI Interaction

Wood (2014) states that first impressions are cognitive processes through which people form judgments quickly, usually within five minutes of encountering someone or experiencing something for the first time. Thus, first impressions are connected to either a negative or a positive experience that consequently strengthens or weakens students' expectations about the performance of the AI-based assessment. These first impressions further affect users' forged preferences (Gronier, 2016; Nourani et al., 2020), and we thus expect that first impressions of new educational AI tools also affect students' learning experiences. This is aligned with the EDT theory (Oliver, 1977; 1980; 1993) that explains users' post-adoption experiences where they evaluate the performance against their prior expectations of the technological system. Hence, the post-adoption experience affects their further willingness to prefer AI-based assessment over the teachers' feedback (cf. Nourani et al., 2020).

Understanding users' first impressions triggered by the use of a new AI-enhanced learning solution helps to plan guidance and support material and teaching methods. These first impressions are considered crucial human processes that describe how users perceive and process interactions and form positive or negative judgments about an object (Alamäki, et al., 2021; Gronier, 2016; Nourani et al., 2020). Affective feelings play an important role in forming these first impressions. Additionally, affective experiences often create stronger media effects than cognitive information alone (e.g., Filippou et al., 2018; Park et al., 2017). Affective experiences and cognitive processing, however, are not separate processes, but rather interrelated when forming first impressions from digital solutions like AI-enhanced learning and teaching systems. To this end, cognitive appraisal theory (Lazarus, 1991) explains that humans cognitively observe and interpret informational elements of objects, and those outcomes affect their initial impression and especially affective experiences. Thus, cognitive observations create information that triggers affective reactions in humans.

Humans start to observe pieces of information from the very first moments of an interaction, and they use this information to form first impressions (Gronier, 2016). Hence, humans form an overall impression of the object or phenomenon before they start to observe and evaluate its details (Kim & Fesenmaier, 2008; Liu et al., 2010). In other words, they begin to form first impressions from several perceived elements of an object right after they have entered a new digital environment (Kim & Fesenmaier, 2018). Consequently, users have sometimes formed their first impression even before they are able to click any button on the interface of a software system, as there are various digital touch points where users interact with service providers (Hallikainen et al., 2019). In addition, there are also different media elements within those digital touchpoints, i.e. software systems, which affect the perceived user experiences. First impressions also affect users' willingness to continue the use of software applications. To this end, Nourani et al. (2020) found that expert users with negative first impressions lose trust in the system and are thus often not willing to use the system later. Hence, first impressions play an important role in the adoption of new technological solutions.

2.3 Expectation Disconfirmation Theory

Expectation Disconfirmation Theory (EDT) aims to understand first impressions in post-adoption situations (Oliver, 1977; 1980, 1993). It has four key concepts: expectations, perceived performance, and the disconfirmation of expectations, all of which influence user satisfaction (Bhattacharjee and Premkumar, 2004; Oliver, 1977; 1980, 1993). In this, expectation refers to the user's attitudes, assumptions, or presumptions before they use a new technological solution. Perceived performance is associated with the actual value the user gains when adopting technological solutions. Disconfirmation of expectations or beliefs is related to the activities where users compare their actual experiences with the original expectations in their minds.

The EDT clarifies the cognitive processes through which students determine their level of satisfaction based on their expectations (Schwarz & Zhu, 2015). It is also related to first impressions (Wood, 2014), as individuals form their initial expectations about a new product or service when they first receive information about it (Schwarz & Zhu, 2015). Research on post-adoption behavior thereby focuses on two areas, namely continued usage of information services as well as their discontinuance (Fan & Suh, 2014). User experience with a new product or service may therefore be both negative and positive, and potentially even during the same usage situation. EDT furthermore shows that users have expectations about the solution they are going to use, and that the actual usage experience provides experiences about the performance of the solution while using it (Alamäki, et al., 2020). This leads to post-purchase satisfaction, which is affected by negative and positive disconfirmation between expectations and performance (Fan & Suh, 2014).

According to EDT, positive user experiences result from the disconfirmation of expectations and the solution's perceived performance. However, users' initial expectations affect the level of user experience as they are the lenses through which the user views their actual experiences. EDT does not make a difference between affective and cognitive factors, which steer the users' expectations and reflection on usage experiences. Cognitive appraisal theory (Lazarus, 1991), on the other hand, explains that users

cognitively observe and interpret information, which creates interpretations, feelings, or experiences that trigger affective reactions in users. Thus, users gain both cognitive information and affective feelings while using a new technological solution or system, and both relate to the expectations and perceived performance of the used solution or system.

3. Materials and Methods

3.1 Research Design and Study Context

This study adopted an exploratory computational text-analysis design implemented within an authentic classroom experiment in higher education. The purpose was to identify and interpret the user-experience issues that students spontaneously reported immediately after using an AI-based essay-grading system. The empirical material consisted of open-ended responses collected after the learning activity. Computational methods were used to identify semantic patterns and group similar responses, while qualitative interpretation was used to examine and explain the resulting clusters.

The classroom experiment was conducted in higher education settings in Austria and Finland. It was designed as an authentic learning activity rather than a controlled laboratory experiment. The study therefore did not aim to establish causal relationships between system use and specific learning outcomes. Instead, it explored students' immediate experiences following their first interaction with an AI-based assessment and essay-grading system in a real-life educational context.

An exploratory design was considered appropriate because the use of advanced AI-based essay assessment in authentic higher education settings remains relatively new. Rather than analysing the responses using predefined categories, the study applied semantic clustering to identify groups of responses with similar meanings. The resulting clusters were subsequently interpreted in relation to research on initial user experiences, trust, feedback, and AI-based assessment.

During the learning activity, students engaged with a topic concerning data-driven business opportunities. A commercial AI-based essay-grading system was used through a customised user interface, and its scoring mechanism had been configured to correspond with the content and learning objectives of the assignment. Students first read a section of a scientific article discussing the transformative effects of cognitive systems on business practices and the opportunities associated with these technologies. They then responded to an assignment prompt by writing a short essay.

After submitting their essays, the students received automated textual feedback and a numerical score generated by the system. They were also shown a model answer illustrating the key elements expected in a high-quality response. Immediately after completing the activity, the students answered open-ended questions concerning their experiences with the AI-based essay-grading system. These post-interaction responses formed the qualitative dataset analysed in the present study.

3.2 Participants and Dataset

Our goal was to investigate a setting in which students use an AI-based essay grading system in a real-life learning situation. As for the collected dataset, we included demographic variables such as gender, age, and educational program, along with students' responses expressing their opinions and overall sentiments toward AI-based essay grading systems. While the survey also contained a variety of other numerical responses from participants, this study only considered open-ended answers to the following question:

"Please, describe your instant feelings related to the usage experiences of this AI-based essay tool".

A total of $n = 109$ responses of varying lengths were collected. Participants' demographics are depicted in Fig. 1. The dataset shows an almost balanced gender distribution and a clear concentration on the 21–30 age group, making it ideal for studying the effectiveness of an AI-based essay grading system, as this age group is likely to be familiar with digital tools and technologies. Furthermore, focusing on business, IT, and other programs situated in the digital field further ensures that the survey captures a broad range of perspectives from individuals with varying levels of exposure to and expectations from AI technologies in academic settings.

3.3 Data Analysis

The open-ended responses were analysed using an exploratory computational text-analysis pipeline consisting of LLM-assisted condensation, semantic embedding, dimensionality reduction, unsupervised clustering, cluster interpretation, and sentiment classification.

The responses of participants varied significantly in length, ranging from brief, single-word reactions such as *"Interesting"* to more elaborate reflections that spanned several sentences. As the diversity in response length could affect clustering (longer responses could dominate the thematic categorization and thus potentially skew the results), we employed a large language model (LLM), OpenAI's GPT-5, to reduce the length of each response while preserving its core meaning. These more concise responses were then compiled and matched with the original responses in a structured format, enabling further analysis and clustering. To ensure the accuracy of the summarization process, the summarized responses were compared with the original inputs to verify that each response had been accurately summarized without losing critical information.

For clustering of survey responses, the first step involved computing embeddings for each concise response. Embeddings are numerical representations of text that capture the semantic meaning of the input. For this purpose, we employed an embedding model (OpenAI's text-embedding-3-small) for generating vector representations of text data. After obtaining the embeddings, their dimensionality was reduced through the UMAP (Uniform Manifold Approximation and Projection) method to make them more manageable and suitable for clustering.

Next, we applied K-means clustering to the embeddings generated from the concise survey responses to identify meaningful groupings based on their semantic similarities. For this purpose, we first found the

optimal number of clusters by evaluating the quality of different numbers of clusters generated by different rounds of clustering through K-means. We iteratively applied K-means clustering to the data, varying the number of clusters from 2 to 30. For each clustering configuration, we computed the Silhouette score, a metric that quantifies how similar an object is to its own cluster compared to other clusters. Higher Silhouette scores indicate more distinct and well-defined clusters.

Figure 4 shows the Silhouette scores for different numbers of clusters. It can be seen that the Silhouette score peaked when the number of clusters was set to 7, indicating that this configuration provided the most distinct and well-defined clustering of the data.

Once the K-means algorithm grouped semantically similar responses into clusters, the subsequent task was to assign meaningful labels that reflected the thematic content of each cluster. To address this, we utilized an LLM to automatically generate appropriate labels for the clusters based on the content of the responses within them.

After the clustering and labeling of survey responses, we employed an LLM to automatically assign sentiments to each response within the clusters. The model was tasked with evaluating the sentiment of the response in the context of the thematic label assigned to its cluster. The final dataset, comprising original responses, concise responses, cluster labels, cluster explanations, and associated sentiments, was used for analysis.

Although the initial grouping was computationally generated, the resulting clusters were interpreted as qualitative themes representing different dimensions of students' immediate experiences.

4. Results

The exploratory clustering analysis identified seven groups of semantically related responses. These clusters provide a data-driven overview of the principal issues students raised immediately after using the AI-based essay-grading system. The clusters of student responses into distinct themes reveal both significant concerns and areas of appreciation. The predominant negative sentiments in the "Evaluation Accuracy Concerns" cluster underscore students' mistrust in the AI system's ability to accurately assess their work, particularly in recognizing the nuanced elements of human writing. This suggests a critical need for improving the accuracy and transparency of AI grading systems to build user confidence.

Conversely, the "Feedback Quality and Improvement" and "User Experience and Convenience" clusters show a more favorable reception, with students acknowledging the usefulness and ease of the AI tool in enhancing their writing skills and offering a convenient, user-friendly experience. However, the "Need for Human Touch in Feedback" cluster reflects a strong desire among students for more personalized feedback, indicating that while AI can be a valuable supplementary tool, it may not be able to fully replace the human element in educational assessments.

The mixed sentiments in the "Mixed Opinions on AI Grading Effectiveness" and "User Trust and Skepticism" clusters further highlight the divided views on the effectiveness and trustworthiness of AI systems. These findings suggest that while there is potential for AI in educational contexts, its implementation must be carefully managed to address concerns related to accuracy, transparency, and the preservation of human interaction.

Table 1 presents selected examples of the original student responses and their corresponding concise versions generated by the LLM. The examples illustrate how responses of different lengths were reduced while preserving their central meaning. The table includes only a sample of the dataset; the same summarization procedure was applied to all 109 responses before the subsequent embedding and clustering stages.

Table 2 presents the seven cluster labels and their descriptions, together with selected examples of the concise responses assigned to each cluster. Table 3 provides selected examples illustrating the complete analytical output for individual responses. For each example, the table shows the original student response, its LLM-generated concise version, the assigned cluster, and the sentiment classification. These examples were selected from different clusters to demonstrate how the analytical process was applied and do not constitute the complete dataset.

Figure 3 shows that students review the cluster " Evaluation Accuracy Concerns " most negatively. This cluster points out that students have concerns regarding the accuracy of AI-based assessment. Additionally, human touch in assessment is expected since the comments related to "Need for Human Touch in Feedback" cluster occurred. However, they feel that the overall User Experience and Convenience cluster is positive without any negative comments. Similarly, "Feedback Quality and Improvement" and "Overall Experience with the AI Tool" clusters collected more positive than negative comments.

5. Discussion

5.1 Theoretical Implications

This study contributes to the ongoing discourse on AI-driven transformation in higher education in general, and more specifically to user experiences of sophisticated AI-based assessment and grading, which are expected to become a critical component of AI-enhanced education. The latest AI-based assessment tools are able to evaluate ill-structured and complex essays that demonstrate higher-order thinking. Thus, this study addresses a new challenge: AI is now capable of performing increasingly complex tasks that have traditionally been carried out by teachers.

The cluster analysis of students' responses to their experiences and expectations of a complex AI-based assessment and grading system reveals a diverse range of sentiments, clustered into distinct thematic areas. These sentiments provide valuable insights into the students' perceptions of AI-based

assessment, highlighting both areas of concern and appreciation. Drawing on the literature of first impressions and experiences (Alamäki, et al., 2021; Gronier, 2016; Nourani et al., 2020) and Expectation Disconfirmation Theory (Oliver, 1980, 1993), we examined students' initial user experiences with AI-based assessment and essay grading in a post-adoption context. In this phase of this study, higher education students had recently engaged with the AI-based assessment and grading system in authentic classroom settings.

The results of the cluster analysis indicate that students generally perceive AI-based assessment positively; however, they continue to express significant concerns regarding its accuracy and the lack of human interaction and oversight. Our findings indicate that, while students appreciate the convenience and usability of AI-based assessment, they express concerns about its fairness in evaluation and grading. Moreover, the desire for more personalized and empathetic feedback suggests that, although AI is perceived as an efficient and valuable assessment tool, it lacks the depth and nuance provided by human evaluators. Thus, our findings support those of Deepshikha (2026), who found that AI-based feedback provides personalized, convenient, real-time feedback and enriches learning experiences. Similarly, Deepshikha's (2026) findings are consistent with our study in showing that concerns about accuracy and algorithmic bias remain and that human oversight is still required to ensure fairness and contextual accuracy. In the following, each cluster is discussed in detail, with an analysis of the associated sentiments and their implications for the development and deployment of AI-based learning and teaching in higher education. These clusters indicate that students evaluate their experiences across multiple dimensions, each emphasizing different thematic concerns.

Evaluation Accuracy Concerns

The "Evaluation Accuracy Concerns" cluster is predominantly associated with negative sentiments. This overwhelming negative sentiment suggests that students are particularly concerned about the accuracy and reliability of AI in grading their essays. This concern is consistent with findings in the literature where the reliability of AI systems in critical evaluative tasks, such as grading, has been questioned. For instance, Perelman *et al.* (2020) argue that AI grading systems may fail to capture the nuances of human writing, leading to inaccurate assessments. The absence of positive sentiments in this cluster further emphasizes the skepticism students have regarding the AI's ability to fairly and accurately evaluate their work. Such may stem, on the one hand, from a perceived lack of grading transparency, and on the other hand, from the fear that the AI might not understand context, creativity, or the subtleties of language as good as human graders do.

Feedback Quality and Improvement

In contrast, the "Feedback Quality and Improvement" cluster is primarily associated with positive sentiments. This indicates that a significant portion of students appreciated the feedback provided by the AI system, viewing it as helpful for improving their writing skills. Positive sentiment in this area suggests that AI was perceived as offering constructive and actionable feedback. The presence of some neutral and negative sentiments, however, indicates that while many students found the feedback useful,

there were still concerns or ambiguities regarding its quality or applicability. These mixed reactions could be due to variability in the feedback's specificity or the perceived relevance to individual student needs.

Mixed Opinions on AI Grading Effectiveness

The “Mixed Opinions on AI Grading Effectiveness” cluster displays a balanced distribution of sentiments, reflecting diverse student views on the effectiveness of AI in grading. This distribution suggests that students are divided regarding the perceived effectiveness of the AI system, with some of them recognizing its potential benefits, and others remaining rather skeptical. The mixed sentiments might reflect the varying levels of familiarity and comfort with AI technologies among students. As noted by Luckin et al. (2016), students' prior experiences and beliefs about technology can significantly influence their perceptions of its effectiveness. The presence of neutral sentiments further suggests that some students are still undecided or require more experience with the system to form a definitive opinion.

Need for Human Touch in Feedback

The “Need for Human Touch” *in Feedback* cluster is predominantly characterized by neutral and negative sentiments, indicating a strong preference among students for human involvement in the feedback process. The absence of positive sentiments here highlights a critical concern, i.e., students might feel that AI, while capable of providing feedback, lacks the empathetic and nuanced understanding that human graders can offer. This aligns with the findings of Ryan and Deci (2000), who suggest that human interaction and feedback are crucial for fostering motivation and engagement in learners. The high number of neutral responses may also indicate that while students recognize the potential of AI, they are not yet fully convinced that it can replace the human element in feedback.

Overall Experience with the AI Tool

The “Overall Experience with the AI” Tool cluster shows a more balanced sentiment distribution, indicating a range of experiences among students. The relatively equal distribution of positive and neutral sentiments suggests that while some students had a favorable experience with the AI tool, many others were uncertain. This uncertainty could stem from the novelty of the technology or a learning curve associated with its use. Research by Davis (1989) on technology acceptance models suggests that user experience plays a critical role in the adoption of new technologies. The presence of negative sentiments, albeit lower, indicates that some issues or challenges detracted from the overall experience for certain students, possibly related to usability or the perceived value of the tool.

User Experience and Convenience

The “User Experience and Convenience” cluster is overwhelmingly positive, indicating that students generally find the AI tool easy to use and convenient. This strong positive sentiment is crucial for the adoption of any educational technology, as ease of use is a significant factor in user acceptance (Venkatesh et al., 2003). The minimal presence of neutral sentiment and absence of negative sentiment suggest that the tool's design and functionality met students' expectations in terms of user-friendliness.

and convenience. This positive reception could also be due to the efficiency and time-saving benefits that AI-based systems often provide, which are particularly valued in educational contexts where students manage multiple tasks and deadlines.

User Trust and Skepticism

Finally, the “User Trust and Skepticism” cluster shows a dominance of neutral sentiments, reflecting a cautious approach among students toward trusting the AI system. The high level of neutrality suggests that while students are open to the possibility of AI being trustworthy, they are not yet fully convinced. This skepticism is consistent with broader discussions in the literature on AI and trust, where trust in AI systems is often seen as contingent on transparency, reliability, and the perceived fairness of the system (Siau & Wang, 2018). The presence of negative sentiments further underscores the need for developers to address issues related to trust, possibly through better communication about how the AI works and assurance that the AI's decisions are unbiased and grounded in sound reasoning.

Understanding students’ initial experiences can help educators design more effective AI-enhanced learning and teaching environments. Initial experiences are important, as they influence how individuals begin to evaluate the characteristics of a given object (Kim & Fesenmaier, 2008; Liu et al., 2010). Prior studies suggest that initial experiences constitute crucial cognitive processes that shape how individuals interpret interactions and form positive or negative judgments about new products and services (Alamäki et al., 2021; Gronier, 2016; Nourani et al., 2020). Hence, the findings of this study assist educators and teachers in creating more effective implementation strategies for AI-based assessment and grading tools in higher education. Additionally, the findings of this study highlight the need for human–AI hybridity in assessment practices where orchestrating the roles of AI, teachers, assessment formats, and learning processes becomes more important. Thus, teachers should balance automated feedback with human oversight when designing implementation strategies for AI-based assessment in both formative and summative contexts.

5.2 Practical Implications

The identified seven clusters provide a novel framework for the design and evaluation of AI-enhanced assessment and essay grading practices in higher education. From the perspective of educators and teachers, our study also highlights the importance of a balanced approach to integrating AI technologies into educational environments. While AI offers several benefits in terms of efficiency and feedback quality, the human element remains essential, particularly in areas requiring nuanced judgment and empathy. Additionally, AI-based assessment and grading of learning tasks, such as essays, support students’ reflective learning process at home or during remote learning, where 24/7 feedback plays a crucial role in learning. If the AI-based assessment is used for formal assessment and grading of the course, not just for supporting the learning process, students’ concerns related to evaluation accuracy should be managed with careful system testing with specific learning objectives and materials.

6 Limitations and Future Research

A key limitation of this study is the relatively small sample size, which focused exclusively on higher education students. This restricts the applicability of the results to other educational sectors. Additionally, the AI-based grading system used in this exploratory study may have device-related limitations that affect the generalizability of findings. Although large language model (LLM)-based chatbots have rapidly gained popularity among higher education students, their hands-on experience with AI-enhanced assessment tools remains limited. Consequently, students' perceptions of grading accuracy may have been influenced by their broader expectations of LLM-based AI chatbots rather than by extensive practical use of this AI-based essay grading system.

Nevertheless, the relationship between students' expectations and their experiences with AI-based assessment tools warrants further investigation. Additional research is needed to explore how AI-based grading and feedback mechanisms can support actual learning, particularly in online and self-study courses where teacher support is minimal or absent. Finally, the clusters identified in this study suggest potential as a framework for evaluating AI-based assessment and essay grading in higher education. Future research should examine more deeply the validity of the framework and whether its clusters can be systematically applied to guide the development and evaluation of AI-enhanced assessment practices. Further studies should also focus on enhancing the reliability of AI grading systems, fostering greater trust through transparency, and exploring hybrid models where AI and human evaluators work in tandem to optimize educational outcomes.

The computational clustering procedure was exploratory and should not be treated as equivalent to a fully validated qualitative coding framework. Cluster membership may have been influenced by the LLM-based condensation of responses, embedding-model characteristics, dimensionality-reduction parameters, and K-means initialization. Further studies should validate the thematic structure through independent human coding, alternative clustering configurations, and additional datasets.

Declarations

Consent to Publish:

Not applicable.

Consent to participate:

Electronic informed consent to participate was obtained from all participants before data collection. At the beginning of the online questionnaire, participants were informed about the study and asked whether they consented to participate. Only participants who selected "yes" were given access to the questionnaire; selecting "no" closed the questionnaire. Participation was voluntary, and the responses used in the study were processed in anonymized form.

Ethical Approval and accordance: The study was reviewed and approved by the MCI Ethics Committee of Management Center Innsbruck, Austria, on 15 December 2022 (reference number 20221203). The

research was conducted in accordance with the ethical guidelines and requirements of the MCI Ethics Committee and the Finnish National Board on Research Integrity (TENK).

Competing interests:

The authors declare no competing interests.

Clinical trial number

Not applicable.

Funding statement:

This study was funded by the AI Driver project, financed by the Ministry of Education and Culture, Finland, under grant number OKM/108/523/2021.

Author Contribution

A.A and S.S prepared the study, conducted the survey, and wrote the Introduction, Related Work, and Discussion sections. U.A.K performed the machine learning-based data analysis and wrote the Materials and Methods, and Results sections.

Data Availability

The dataset generated and analyzed during this study is publicly available in the Zenodo repository under DOI: <https://doi.org/10.5281/zenodo.21396766>.

References

1. Alfaleh M. Sustainable AI-Driven Assessment in Higher Education: A Systematic Review of Fairness, Transparency, Pedagogical Innovation, and Governance. *Sustainability*. 2026;18(2):785.
2. Alamäki A, Dirin A, Suomala J. Students' expectations and social media sharing in adopting augmented reality. *Int J Inform Learn Technol*. 2020;38(2):196–208.
3. Alamäki AV, Dirin A, Suomala J, Rhee C. Students' Experiences of 2D and 360° Videos With or Without a Low-Cost VR Headset: An Experimental Study in Higher Education. *J Inform Technol Educ Res*. 2021;20:309–29.
4. Alamäki A, Khan UA, Kauttonen J, Schlögl S. An experiment of AI-based assessment: Perspectives of learning preferences, benefits, intention, technology affinity, and trust. *Educ Sci*. 2024;14(12):1386.

5. Bhattacharjee A, Premkumar G. Understanding changes in belief and attitude toward information technology usage: a theoretical model and longitudinal test. *MIS Q.* 2004;28(2):229–54.
6. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 1989;13(3):319–40.
7. Deepshikha D. A systematic review on the future of educational assessment: AI-driven grading and personalised feedback in higher education. *Artif Intell Educ.* 2026;2(2):75–115.
8. Dirin A, Laine T, Alamäki A. Managing Emotional Requirements in a Context-Aware Mobile Application for Tourists. *Int J Interact Mob Technol.* 2018;12(2):177–96.
9. Fan L, Suh YH. Why do users switch to a disruptive technology? An empirical study based on expectation-disconfirmation theory. *Inf Manag.* 2014;51(2):240–8.
10. Filippou J, Cheong C, Cheong F. A model to investigate preference for the use of gamification in a learning activity. *Australasian J Inform Syst.* 2018;22:1–23.
11. Gao R, Merzdorf HE, Anwar S, Hipwell MC, Srinivasa A. Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*; 2024. p. 100206.
12. Gronier G. Measuring the first impression: Testing the validity of the 5-second test. *J Usability Stud.* 2016;12(1):8–25.
13. Hallikainen H, Alamäki A, Laukkanen T. Individual preferences of digital touchpoints: A latent class analysis. *J Retailing Consumer Serv.* 2019;50:386–93.
14. Jahanmir SF, Cavadas J. Factors affecting late adoption of digital innovations. *J Bus Res.* 2018;88:337–43.
15. Kim H, Fesenmaier DR. Persuasive design of destination websites: An analysis of first impression. *J Travel Res.* 2008;47(1):3–13.
16. Knutson B, Greer SM. Anticipatory affect: Neural correlates and consequences for choice. *Philosophical Trans Royal Soc B: Biol Sci.* 2008;363(1511):3771–86.
17. Liu C, White RW, Dumais S. (2010). Understanding web browsing behaviors through Weibull analysis of dwell time. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland*, 379–386.
18. Luckin R, Holmes W, Griffiths M, Forcier LB. *Intelligence Unleashed: An argument for AI in education.* Pearson. University College London; 2016.
19. Luo J, Zheng C, Yin J, Teo HH. Design and assessment of AI-based learning tools in higher education: A systematic review. *Int J Educational Technol High Educ.* 2025;22(1):42.
20. Minn S. AI-assisted knowledge assessment techniques for adaptive learning environments. *Computers Education: Artif Intell.* 2022;3:100050.
21. Nourani M, King J, Ragan E. (2020). The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (Vol. 8, pp. 112–121).*

22. Oliver RL. Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *J Appl Psychol*. 1977;62(4):480.

23. Oliver RL. A cognitive model of the antecedents and consequences of satisfaction decisions. *J Mark Res*. 1980;17(4):460–9.

24. Oliver RL. Cognitive, affective, and attribute bases of the satisfaction response. *J Consum Res*. 1993;20(3):418–30.

25. Park B, Blevins E, Knutson B, Tsai JL. Neurocultural evidence that ideal affect match promotes giving. *Soc Cognit Affect Neurosci*. 2017;12(7):1083–96.

26. Perelman L. (2020). The BABEL generator and e-Rater: 21st Century writing constructs and automated essay scoring (AES). *J Writ Assess* 13.1.

27. Ramesh D, Sanampudi SK. An automated essay scoring system: a systematic literature review. *Artif Intell Rev*. 2022;55(3):2495–527.

28. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am Psychol*. 2000;55(1):68–78.

29. Schwarz C, Zhu Z. The impact of student expectations in using instructional tools on student engagement: A look through the expectation disconfirmation theory lens. *J Inform Syst Educ*. 2015;26(1):47–58.

30. Siau K, Wang W. Building trust in artificial intelligence, machine learning, and robotics. *Cut Bus Technol J*. 2018;31(2):47–53.

31. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: Toward a unified view. *MIS Q*. 2003;27(3):425–78.

32. Wood TJ. Exploring the role of first impressions in rater-based assessments. *Adv Health Sci Educ*. 2014;19(3):409–27.

Tables

Table 1: Illustrative examples of original student responses and their corresponding LLM-generated concise responses

Original student response	LLM-generated concise response
<i>This may be the future.</i>	Future potential
<i>I have mixed feelings. I think the grading wasn't right but I also think it's efficient.</i>	Mixed feelings on efficiency and accuracy
<i>I thought it was cool but it feels kind of cold because there is not an actual human giving me results or feedback.</i>	Cool but lacks human touch
<i>I don't feel like the AI tool could give an exact feedback from the essay. It seems like it only recognizes the words but not if the point is understood correctly.</i>	Limited understanding of essay content
<i>When the comment on your essay comes by technology it does not feel personal.</i>	Impersonal feedback
<i>The essay tool worked well but I would prefer teacher to check my essays if I had to choose.</i>	Prefer human grading
<i>It was very easy to use and the feedback was accurate.</i>	Easy to use, accurate feedback
<i>The design and confusion of the test does not inspire confidence in artificial intelligence. I assumed artificial intelligence would be a little more effective in finding meanings in the text, but I was disappointed with the result.</i>	Disappointed in AI's effectiveness
<i>Interested.</i>	Interest in AI grading
<i>The speed of the test results and the feedback surprised me. Also, the ease of the test and how precisely the topic was defined and how the result came out of these was amazing.</i>	Surprised by speed and precision
<i>I can agree with some of the feedback that I got, but I think, that the AI-tool didn't evaluate my conclusions so well, considering the article.</i>	Inaccurate evaluation of conclusions
<i>I'm not sure if I can trust the tool's evaluation completely but it looks promising.</i>	Promising but not fully trustworthy
<i>nice that the feedback is given immediately.</i>	Immediate feedback appreciated
⋮	⋮

Table 2: Seven cluster labels, their descriptions, and illustrative examples of concise responses

Cluster label	Cluster description	Concise responses
Mixed Opinions on AI Grading Effectiveness	This cluster captures the varied opinions about the effectiveness of the AI grading system, with some respondents expressing disappointment and others acknowledging its potential benefits for learning.	['prefer human grading', 'disappointed in ai's effectiveness', 'interest in ai grading', 'beneficial for learning, but lacks for actual grading', 'nice tool to have', 'curious to see grade, fine using the tool', 'disappointing layout and usage', 'effective but limits creativity', 'doubts ai's understanding of sentences', 'inaccurate and unhelpful ai tool', 'interesting to use ai tools', 'questioning keyword search vs. deeper analysis', 'ai tool aids in organizing thoughts and improving writing']
Overall Experience with the AI Tool	This cluster summarizes the general sentiments about the experience of using the AI grading system, reflecting a mix of enjoyment, interest, and some feelings of robotic interaction.	['enjoyable experience', 'interesting experience', 'interesting', 'satisfied', 'exciting', 'somewhat robotic', 'decent but hard to understand', 'creepy', 'interested', 'not so good', 'not good', 'good', 'fun']
User Experience and Convenience	This cluster highlights the positive aspects of the AI grading system, focusing on its user-friendliness, efficiency, and the convenience it offers to users.	['easy to use, accurate feedback', 'user-friendly and reliable', 'quick and effective answers', 'easy to use and pleasant', 'easy to use, accuracy concerns', 'boosts productivity', 'quick and easy process', 'accurate and convenient data analysis', 'easy to use', 'useful for teaching', 'useful for quick answers, no follow-up questions', 'useful for learning', 'could increase productivity', 'practical, time-saving, efficient, good for ideas', 'convenience and time-saving benefits']
User Trust and Skepticism	This cluster reflects the respondents' feelings of uncertainty and skepticism regarding the AI grading system's reliability and accuracy, highlighting their mixed emotions and trust issues.	['future potential', 'surprised by speed and precision', 'promising but not fully trustworthy', 'confused and mistrusting', 'skeptical about results, first-time user', 'skeptical about accuracy', 'trust issues, surprised by quality of answer', 'skeptical', 'hesitant and slightly nervous', 'unsure', 'confusion', 'astonished']
Need for Human Touch in Feedback	This cluster emphasizes the desire for more personalized and human-like feedback, indicating that while the AI is seen as a useful tool, it lacks the depth and understanding that human evaluators provide.	['cool but lacks human touch', 'impersonal feedback', 'interesting tool, desired more personalized feedback', 'less developed than chatgpt, prefers self-checking', 'great instant feedback, lacks personalization and follow-up options', 'good base, needs human review', 'fast but lacks human feedback', 'quick results, no personal touch', 'interesting but prefer personal feedback', 'great tool, but should not replace human feedback', 'needs human touch for evaluation and understanding', 'accurate feedback, prefer human evaluation', 'trust in ai for feedback, prefer professor for final', 'effective but lacks professor's expertise', 'unpersonal but fair', 'fast but lacks depth', 'prefer human feedback for personal touch', 'evaluation criteria needed', 'easier to write academically']

Feedback Quality and Improvement	This cluster deals with the quality of feedback provided by the AI, with respondents expressing a range of experiences from appreciation for immediate feedback to a desire for more detailed and constructive critiques.	['immediate feedback appreciated', 'neutral feedback', 'unrelated comment', 'immediate feedback, surprising', 'average feedback, wanted detailed explanations', 'positive feedback', 'detailed improvement information provided', 'interesting corrections', 'expect detailed feedback on faults', 'neutral, not very useful feedback', 'great feedback, improved response quality', 'uncertainty about feedback justification', 'needs improvement', 'relieved with constructive feedback', 'surprising and validating', 'needs deeper feedback', 'helpful feedback provided']
Evaluation Accuracy Concerns	This cluster addresses the respondents' concerns about the accuracy and depth of the AI's evaluations, including issues with misleading results and the AI's understanding of essay content.	['mixed feelings on efficiency and accuracy', 'limited understanding of essay content', 'inaccurate evaluation of conclusions', 'good recognition, example essay too long', 'misleading results based on incomplete reading', 'inaccurate grading for nonsensical text', 'confusing rating, frustrated but would use again', 'insulting feedback for low grade', 'encourages repetitive wording', 'inaccurate analysis, expected lower rating', 'not fully accurate', 'high score feels untrustworthy', 'focus on quantity over quality', 'no sympathy marks', 'relies on prewritten content, not reliable', 'ai lacks depth in analysis']

Table 3: Illustrative examples of original responses, concise responses, cluster assignments, and sentiment classifications

Original student response	LLM-generated concise response	Assigned cluster	Sentiment classification
<i>This tool is effective and time saving. we can get the feedback quickly and accurately. we can avoid the personal bias in exams. But AI tools limit the creativity and critical thinking.</i>	effective but limits creativity	Mixed Opinions on AI Grading Effectiveness	neutral
<i>At first, I felt curious on how it might work but was disappointed quickly because of the layout and usage of the system - it was way less satisfying than I expected from a modern system.</i>	Disappointing layout and usage	Mixed Opinions on AI Grading Effectiveness	negative
:	:	:	:
<i>Still seems a bit robotic like, but not too bad.</i>	Somewhat robotic	Overall Experience with the AI Tool	neutral
<i>Not so good</i>	Not so good	Overall Experience with the AI Tool	negative
<i>This was a fun trial</i>	Enjoyable experience	Overall Experience with the AI Tool	positive
:	:	:	:
<i>It felt rather like a true false question. Although I think it would be great addition to teaching</i>	Useful for teaching	User Experience and Convenience	positive
<i>The usage was easy but I don't think the results were completely accurate.</i>	Easy to use, accuracy concerns	User Experience and Convenience	neutral
:	:	:	:
<i>hesitant & slightly nervous</i>	Hesitant and slightly nervous	User Trust and Skepticism	neutral
<i>I thought the results were too good to be true. I couldn't quite understand how it figured those things out of my own answer.</i>	Skeptical about accuracy	User Trust and Skepticism	negative
:	:	:	:
<i>I tend to prefer humans to correct my work since I feel like they give a more personal and detailed feedback. The AI-based essay tool to me only seemed to check for key words necessary to be included in the text rather</i>	Prefer human feedback for personal touch	Need for Human Touch in Feedback	negative

than understandability and coherence. Furthermore, you only see your achieved percentage but not a grading scale indicating what input is expected for which grade level. And with this tool, there is no possibility to ask questions on the grading afterwards.

Figures

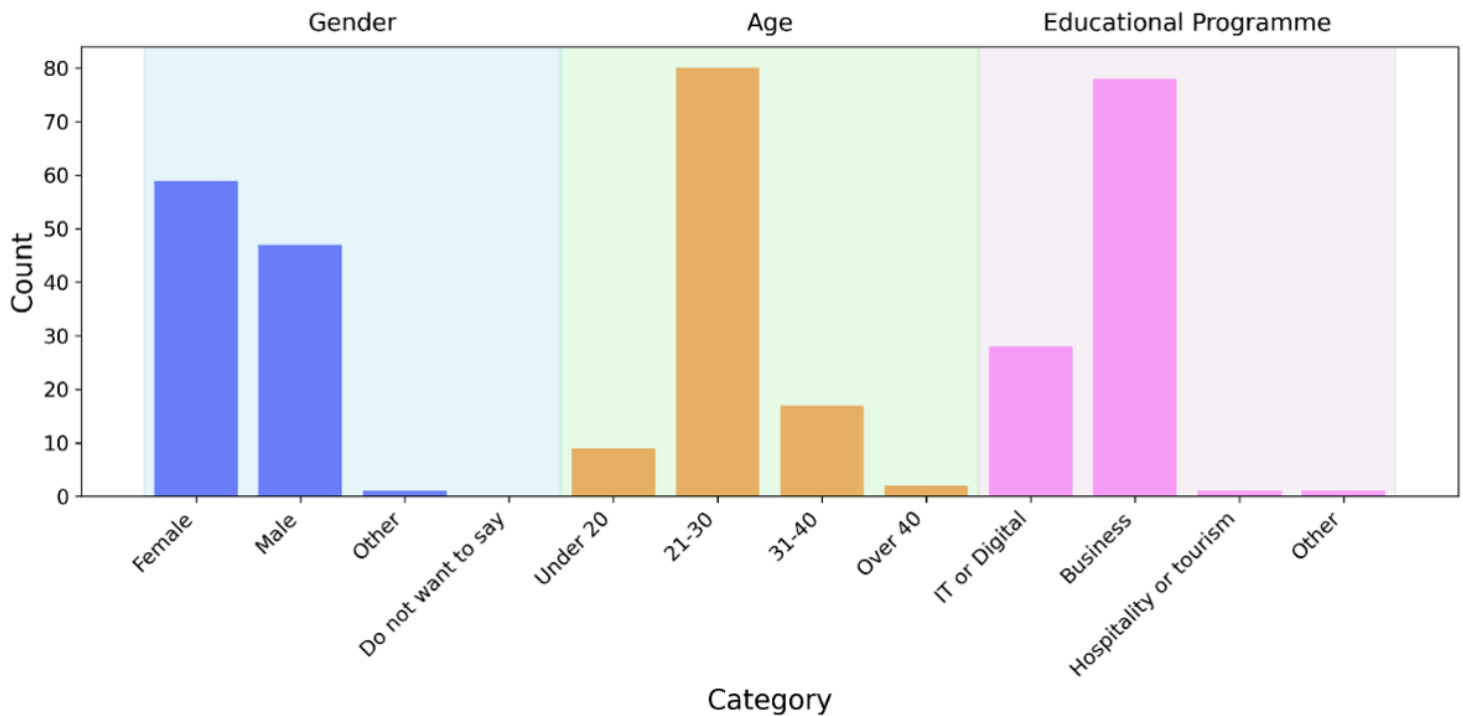


Figure 1

Distribution of participants' demographics

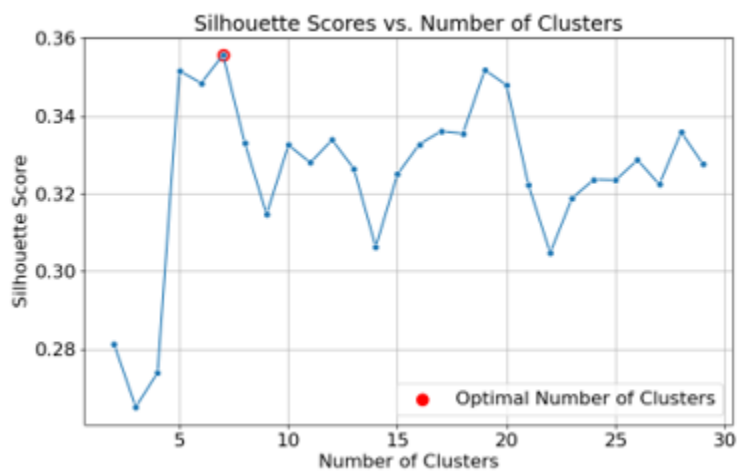


Figure 2

Silhouette Scores vs. Number of Clusters. The red marker highlights the optimal number of clusters (n=7)

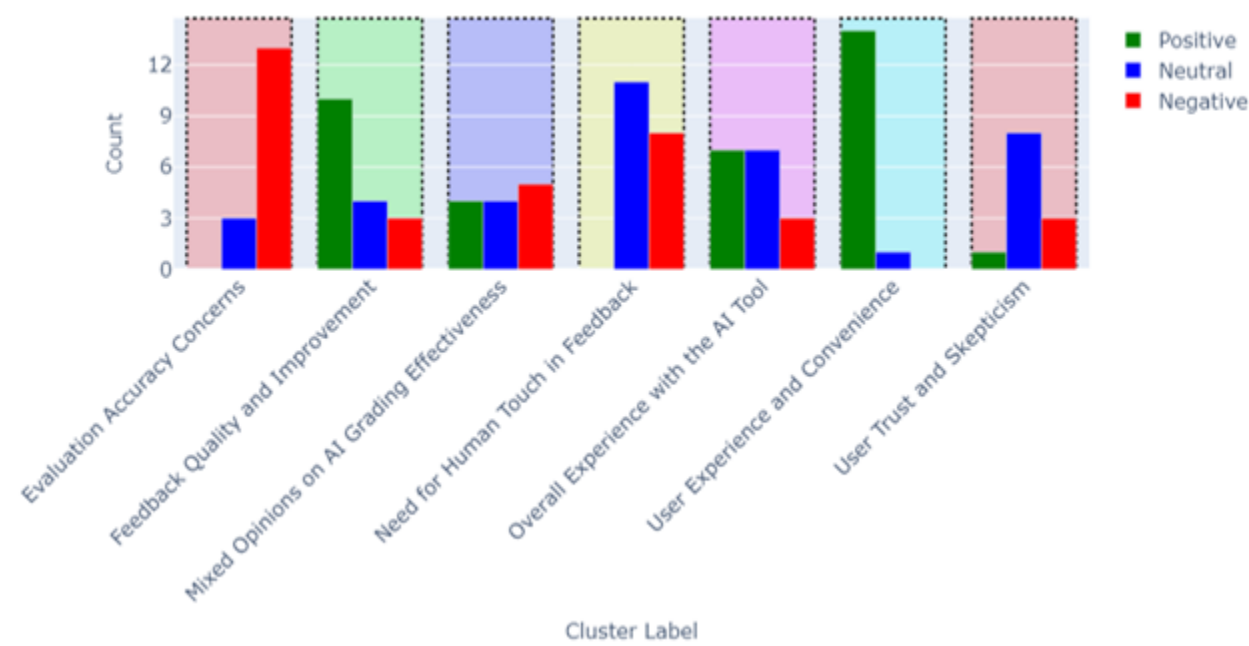


Figure 3

Distribution of sentiments across clusters