Contemporary Educational Technology, 2026, **18**(1), ep623
ISSN: 1309-517X (Online)

OPEN ACCESS

**Review Article**

# Human and AI-generated feedback in higher education: A systematic review of effectiveness and student perceptions

**Candy Haydee Guardia-Paniura** [1]
0000-0003-1141-2820

**Timoteo Cueva-Luza** [2]
0000-0002-2199-5671

**Favio Mauricio Cruz-Carpio** [3]
0000-0001-6605-6734

**Raúl Reynaldo Ito-Díaz** [4]
0000-0003-3370-5990

**David Victor Apaza-Paco** [5]
0000-0002-8553-2879

**Nilda Rosas-Rojas** [6]
0000-0002-6509-5528

**Benedicta Mamani-Mamani** [4]
0000-0003-4502-4981

**Ángel Terrero-Pérez** [7]
0000-0002-7338-3642

**Renato Yassutaka Faria Yaedú** [7]
0000-0002-2576-7887

**Mariela Peralta-Mamani** [7,8*]
0000-0002-0243-9194

[1] National Pedro Ruíz Gallo University, Lambayeque, PERU
[2] National University of Moquegua, Moquegua, PERU
[3] Technological University of Peru, Lima, PERU
[4] National University of Juliaca, Puno, PERU
[5] National University of San Agustín of Arequipa, Arequipa, PERU
[6] Peruvian Union University, Puno, PERU
[7] Hospital for Rehabilitation of Craniofacial Anomalies, University of São Paulo, Bauru, São Paulo, BRAZIL
[8] Faculdade do Centro Oeste Paulista, Piratininga, São Paulo, BRAZIL
[*] Corresponding author: marielaperalta@alumni.usp.br

**Citation:** Guardia-Paniura, C. H., Cueva-Luza, T., Cruz-Carpio, F. M., Ito-Díaz, R. R., Apaza-Paco, D. V., Rosas-Rojas, N., Mamani-Mamani, B., Terrero-Pérez, Á., Yaedú, R. Y. F., & Peralta-Mamani, M. (2026). Human and AI-generated feedback in higher education: A systematic review of effectiveness and student perceptions. *Contemporary Educational Technology, 18*(1), ep623. https://doi.org/10.30935/cedtech/17863

**ARTICLE INFO**

Received: 17 Dec 2024
Accepted: 29 Dec 2025

**ABSTRACT**

This study aims to compare the feedback provided by human professors and ChatGPT on university students' work and to report on students' perceptions of both types of feedback. A systematic review was conducted following PRISMA 2020 guidelines. Databases research included Web of Science, Scopus, EBSCO, ACM Digital Library, and IEEE Xplore, with additional gray literature sources, until October 2024. Inclusion criteria were cross-sectional studies evaluating university students' work, comparing feedback from ChatGPT with human professors.

Data extraction was performed using a standardized form, and risk of bias was assessed with the Joanna Briggs Institute critical appraisal tool. A narrative synthesis of the results was made. PROSPERO registration number: CRD42024566691. This review included 8 studies with 461 students. ChatGPT feedback was detailed and rapid, while human feedback was valued for its personalization and emotional support. Students appreciated the detailed and immediate nature of ChatGPT feedback but noted its lack of emotional nuance and context-specific guidance. Human feedback was preferred for addressing individual learning needs and providing affective support. A combination of both types of feedback to maximize benefits. ChatGPT can assist human teachers by providing detailed and timely feedback to university students. However, human supervision is essential to ensure feedback is nuanced and contextually appropriate. A hybrid approach can optimize the learning experience in higher education. Further research is necessary to explore AI applications in educational settings and understand their impact on learning outcomes.

**Keywords:** higher education, ChatGPT, feedback, systematic review

## INTRODUCTION

Feedback is any information given to a student after their response to inform them about their performance. Educational feedback is an effective approach to enhance student learning. However, it can be labor-intensive, which motivates the use of automated feedback tools (Bauer et al., 2023; Dai et al., 2023a).

Providing feedback to higher education students is an essential skill for teachers and significantly influences the learning process. The development of writing skills in university students is crucial for their academic and professional success. Constructive feedback from teachers plays a fundamental role by offering ideas and recommendations to improve students' writing abilities. This contributes to a deeper understanding and enhances the ability to communicate effectively, having a significant impact on the professionalization of higher education (Al-Bashir et al., 2016; AlGhamdi, 2024).

Providing individualized feedback for the student becomes challenging, as teachers are often overwhelmed in large classes of students. Thus, these challenges have led to looking for innovative solutions, such as automated feedback using artificial intelligence (AI) (Al-Bashir et al., 2016; AlGhamdi, 2024). ChatGPT is an intelligent AI-developed chatbot that was launched in November 2022. It has multiple applications and the ability to generate various forms of text, answer questions, and provide translations (Xiao & Zhi, 2023).

AI is a powerful data analysis tool that enhances the quality of feedback, which can boost productivity. Tools like ChatGPT can be useful for this purpose, providing individualized and timely feedback. However, they are limited in terms of quality, authenticity, and emotional intelligence. People may have a negative perception of these tools (Tong et al., 2021).

In recent years, research on the use of AI in education has expanded considerably. Nevertheless, few studies have directly compared feedback generated by ChatGPT with that provided by human professors, particularly regarding students' perceptions, emotional responses, and learning outcomes. Understanding these differences is essential for guiding evidence-based pedagogical practices and ensuring the ethical and effective integration of AI in higher education. This gap in the literature highlights the need for a systematic synthesis that examines the advantages and limitations of AI- and human-generated feedback, providing insights into how both can be combined to enhance the teaching-learning process. Therefore, this systematic review aims to compile and critically analyze studies comparing feedback generated by ChatGPT with that provided by human professors in higher education. The review was guided by the following research question: Among higher education students, how does feedback generated by ChatGPT compare with that provided by human professors in terms of effectiveness, quality, and students' perceptions?

## MATERIALS AND METHODS

### Protocol and Registration

This systematic review followed the methodological framework proposed by the PRISMA 2020 statement (Page et al., 2021) to ensure transparent and standardized reporting. The study protocol was prospectively

registered in the PROSPERO database (international prospective register of systematic reviews) under the identification number CRD42024566691 (www.crd.york.ac.uk/PROSPERO)

## Eligibility Criteria

### *Inclusion criteria*

The systematic review included studies that met the following PECO strategy:

(1) participants: university students,

(2) exposure: feedback from ChatGPT,

(3) control: feedback from human professors, and

(4) outcome: effectiveness of the feedback and student perceptions of the feedback.

All cross-sectional studies that evaluated the work of university students comparing feedback from ChatGPT with feedback from human professors were included. Both cross-sectional and short-term longitudinal designs were eligible if they involved an evaluation of university students' academic work comparing ChatGPT- and human-generated feedback.

### *Exclusion criteria*

Studies that included postgraduate students or university professors, or that used ChatGPT for purposes other than generating feedback on university work, were excluded. Additionally, reviews and letters to the editor were excluded.

## Exposure and Control

The exposure in this systematic review was feedback provided by ChatGPT, an AI chatbot. The control was feedback given by human professors. This comparison aimed to evaluate the effectiveness and perceptions of both feedback sources in the context of higher education.

## Information sources and Search

The search strategy included the following electronic bibliographic databases: Web of Science, Scopus, EBSCO, ACM Digital Library, and IEEE Xplore. Additionally, gray literature was searched using the Brazilian digital library of theses and dissertations, OpenGrey, ProQuest, and Google Scholar (first 100 records). The search terms related to "higher education," "ChatGPT," and "feedback," and were combined using Boolean operators "OR" and "AND." No restrictions were applied to the year of publication or language initially. Additionally, a manual search was conducted to identify eligible studies.

Studies published up to October 2024 were included in the systematic review. The search strategy was adapted for each database (see **Appendix A**). All collected records were imported into EndNote Web (www.myendnoteweb.com), where duplicates were removed.

## Study Selection

All records were imported into Rayyan software for the initial phase of study selection. In this phase, two reviewers (ÁT-P and MP-M) independently screened the titles and abstracts. During the second phase, the full texts of potentially eligible studies were reviewed to confirm whether they met the eligibility criteria. Any discrepancies between the reviewers were resolved by consensus.

## Data Extraction and Data Items

Data extraction was performed independently by two reviewers (ÁT-P and MP-M) using a standardized form in Microsoft Excel. If there was any discrepancy, it was resolved by consensus. The extracted data included: first author and year of publication, geographic region, total number of participants, participant demographics (age and sex), type of work evaluated, details of the feedback provided (ChatGPT and human professors), study variables and results (**Appendix B**, **Appendix C**, **Appendix D**, and **Appendix E**).

### Risk of Bias in Individual Studies

The risk of bias in individual studies was assessed using the "Joanna Briggs Institute critical appraisal checklist for analytical cross-sectional studies" from the Joanna Briggs Institute (Moola et al., 2020). This tool consists of eight items that evaluate various criteria of the studies, including the clear definition of inclusion criteria, detailed description of the subjects and the study setting, the validity and reliability of exposure measurement, the use of objective standard criteria for condition measurement, identification of confounding factors and strategies to manage them, the validity and reliability of outcome measurement, and the adequacy of the statistical analysis used. The possible responses to each item were: "yes," "no," "unclear," or "not applicable," as appropriate. Two reviewers independently assessed each study, with discrepancies resolved by consensus.

### Summary Measures

Effectiveness was measured using various metrics such as the intraclass coefficient, absolute values, and percentages. Student perceptions were assessed through surveys and Likert scale ratings, capturing measures of satisfaction and perceived quality of the feedback.

### Synthesis of Results

A narrative synthesis of the results was structured around the comparison of feedback from ChatGPT and human professors, including student perceptions. The data synthesis focused on identifying common themes and differences in feedback effectiveness and student perceptions across the included studies.

## RESULTS

### Study Selection

A total of five databases and gray literature were searched, yielding 288 records. After excluding duplicates, 246 records were included for the title and abstract screening phase. Of these, 210 studies were excluded as they did not meet the eligibility criteria. Out of the remaining 36 studies, 16 were deemed ineligible for this review. Therefore, 20 studies proceeded to the full-text review phase.

Thus, 12 studies were excluded for various reasons. Two studies were excluded from being reviews (Bauer et al., 2023; Cowling et al., 2023), one study was excluded for using ChatGPT in the supervision of postgraduate student investigations (Dai et al., 2023b), and one study focused on high school teachers' perception of ChatGPT use (El Sayary, 2023). Another article dealt with ChatGPT feedback without comparing it to human feedback (Yan, 2024). Seven studies were excluded because ChatGPT was used to assist in various tasks: a study on student perception using ChatGPT for Java programming (Haindl & Weinberger, 2024), a study on using ChatGPT to create a founding team within an entrepreneurship course (Hammoda, 2024), a study on using ChatGPT to facilitate the development of educational experiences in Roblox (Hoo & Lee, 2023), a study on the effectiveness of ChatGPT as a tool for developing English learning skills (Muniandy & Selvanathan, 2024), a study on using ChatGPT for learning (benefits, barriers, and possible solutions) (Ngo, 2023), a study about RECaP-GPT, which integrates human action and uses ChatGPT-4 as a feedback teaching support tool (Ossa & Willatt, 2023), and a study on using AI for a comprehensive review of existing film courses and AI-recommended courses (Yang et al., 2023). Thus, this systematic review included 8 studies (AlGhamdi, 2024; Escalante et al., 2023; Guo & Wang, 2023; Ivanovic, 2023; Jukiewicz, 2024; Lu et al., 2024; Tossell et al., 2024; Wang et al., 2024) (**Figure 1**).

### Study Characteristics

The included studies were published between 2022 and 2024, as ChatGPT was launched in November 2022. One study was from Saudi Arabia (AlGhamdi, 2024), two studies were from the USA (Escalante et al., 2023; Tossell et al., 2024), three studies were from China (Guo & Wang, 2023; Lu et al., 2024; Wang et al., 2024), one study was from Montenegro (Ivanovic, 2022), and one study was from Poland (Jukiewicz, 2024).

In total, 461 higher education students were included, 202 were men and 150 women, aged 18 to 36 years. Two studies did not report the gender and age of participants (Jukiewicz, 2024; Wang et al., 2024).

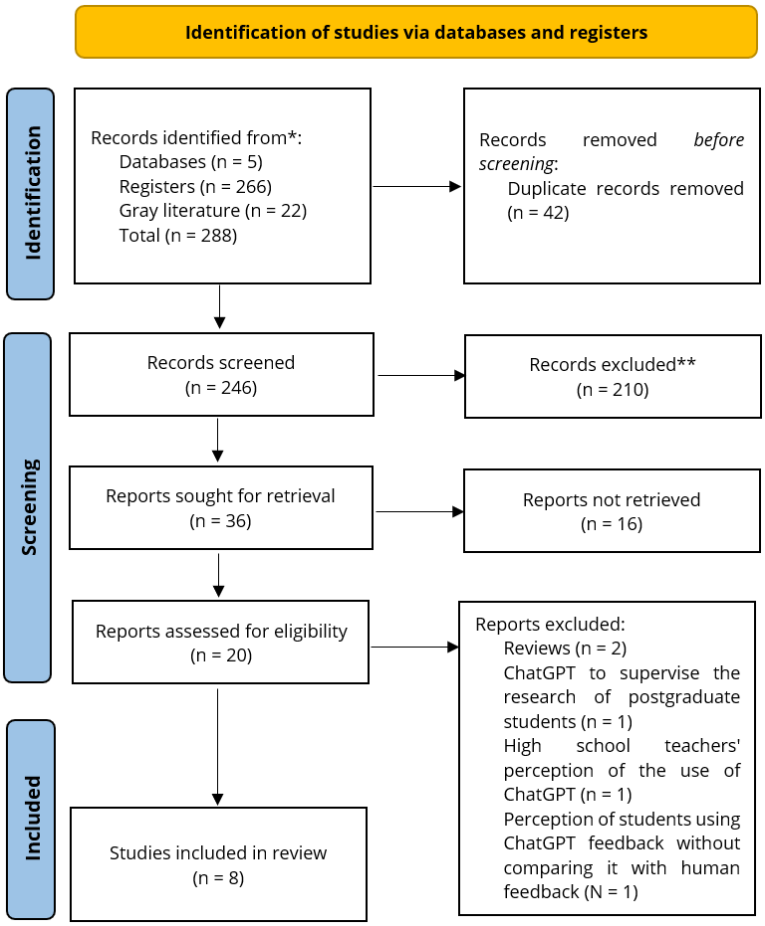


**Figure 1.** Flowchart of study selection for qualitative syntheses (Authors' own elaboration, based on PRISMA 2020 guidelines)

The evaluated work included essay writing as part of semester assignments in various fields and writing assignments in English courses. The analyzed variables included clarity, usefulness, preference, quality, organization, educational value, and confidence in the evaluation. All studies compared feedback from human instructors with feedback from ChatGPT. Two studies used ChatGPT-4 (Escalante et al., 2023; Tossell et al., 2024), four studies used ChatGPT-3.5 (Ivanovic, 2023; Jukiewicz, 2024; Lu et al., 2024; Wang et al., 2024), and two studies did not specify the version used (AlGhamdi, 2024; Guo & Wang, 2023).

Most of the studies evaluated student assignments over a period of 6 weeks (AlGhamdi, 2024; Escalante et al., 2023; Lu et al., 2024), while others evaluated work from a single class session (Guo & Wang, 2023), 8 weeks (Wang et al., 2024), two months (Ivanovic, 2023; Tossell et al., 2024), and 15 weeks (Jukiewicz, 2024) (**Appendix B**, **Appendix C**, **Appendix D**, and **Appendix E**).

## Risk of Bias Within Studies

The quality assessments of the individual studies are listed in **Table 1**. This evaluation is based on the Joanna Briggs Institute critical appraisal tool for analytical cross-sectional studies. The eight included studies exhibit various levels of bias risk. Six studies present a low risk of bias due to their rigorous designs, clear inclusion criteria, and standardized evaluations (AlGhamdi, 2024; Escalante et al., 2023; Guo & Wang, 2023; Ivanović, 2022; Lu et al., 2024; Tossell et al., 2024). However, these studies have limitations that may affect the generalizability of the results. AlGhamdi (2024) included only male students. Tossell et al. (2024) used a small, homogeneous sample of USAFA cadets. Escalante et al. (2023) had self-selected participants and a homogeneous sample. Guo and Wang (2023) worked with a limited sample of five teachers and self-selected participants. Ivanović (2022) faced limitations due to the limited variability in the sample and the possible influence of human evaluators. Lu et al. (2024) presented a limitation in the homogeneity of the sample of Chinese students and the possible influence of the feedback sequence.

**Table 1.** JBI critical appraisal checklist for analytical cross sectional studies

| Questions | AlGhamdi (2024) | Escalante et al. (2023) | Guo and Wang (2023) | Ivanovic (2023) | Jukiewicz (2024) | Lu et al. (2024) | Tossell et al. (2024) | Wang et al. (2024) |
|---|---|---|---|---|---|---|---|---|
| 1. Were the criteria for inclusion in the sample clearly defined? | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 2. Were the study subjects and the setting described in detail? | Yes | Yes | Yes | Yes | No | Yes | Yes | No |
| 3. Was the exposure measured in a valid and reliable way? | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 4. Were objective, standard criteria used for measurement of the condition? | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 5. Were confounding factors identified? | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 6. Were strategies to deal with confounding factors stated? | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 7. Were the outcomes measured in a valid and reliable way? | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| 8. Was appropriate statistical analysis used? | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

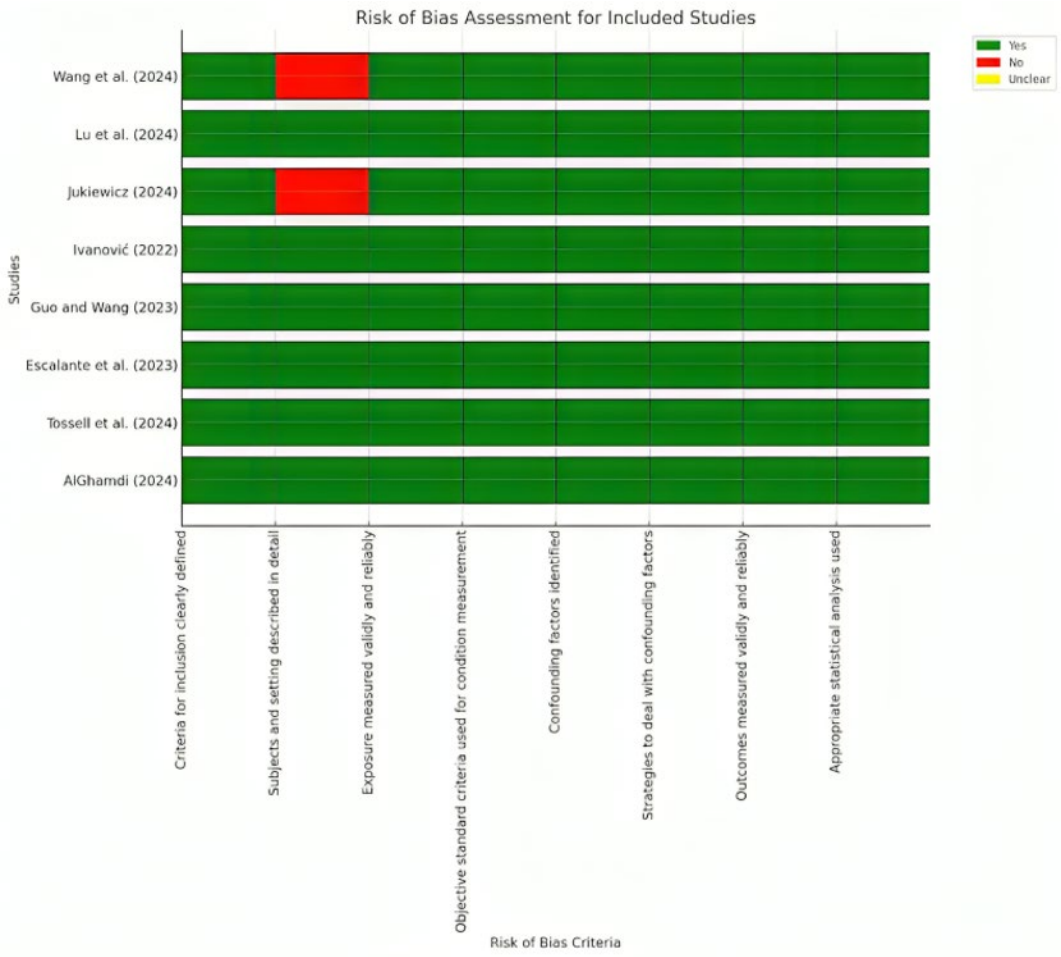


**Figure 2.** Risk of bias graph (Authors' own elaboration using the Joanna Briggs Institute critical appraisal tool)

On the other hand, the studies by Jukiewicz (2024) and Wang et al. (2024) present a moderate risk of bias due to the lack of detailed information about the age and gender of the participants. Although both studies used blind designs and standardized evaluations, the homogeneity of their samples limits the generalizability and representativeness of the results (**Figure 2**).

## Results of Individual Studies

The study conducted by AlGhamdi (2024) found that feedback generated by ChatGPT had emotional, psychological, and educational impacts on first-year computing students. Responses to the feedback ranged from positive emotions, such as motivation and enthusiasm, to negative ones, such as frustration and confusion. Regarding quality and usefulness, some students appreciated the detailed improvements provided by ChatGPT, while others criticized its lack of consistency and personalization. In terms of development and progress, many students acknowledged improvements in their writing skills due to regular and detailed feedback, although some noted the lack of personalization compared to human feedback. In summary, the study highlights the potential of ChatGPT to provide useful and timely feedback but emphasizes the need to complement it with human comments to more effectively address the emotional and educational needs of students.

The study by Escalante et al. (2023) indicates that there were no significant differences between the feedback generated by ChatGPT-4 and human instructors. Approximately the same number of students preferred AI-generated feedback and human feedback. Some characteristics of AI feedback include clarity and specificity, while human feedback is valued for its affective benefits and direct interaction. The results suggest that AI-generated feedback can be incorporated into student essay evaluations without negatively affecting learning outcomes, and they recommend a mixed approach that combines the strengths of both types of feedback.

The results of the study by Guo and Wang (2023) showed that ChatGPT generated longer, more detailed, and specific feedback compared to human instructors, who focused on issues related to content and language. Additionally, ChatGPT provided more balanced comments. Instructors expressed both positive and negative perceptions, noting that ChatGPT can complement their own feedback. However, human supervision and adjustment are necessary to maximize its effectiveness in developing writing skills.

The study by Ivanovic (2023) compared feedback from human instructors and ChatGPT, finding good consistency and reliability (intraclass correlation coefficient [ICC] of 0.8). ChatGPT can evaluate work in less than 30 seconds and provide detailed analysis similar to human instructors, objectively based on training data. It may be more lenient with minor errors, often giving slightly higher grades. However, it is not capable of capturing emotional and cultural nuances and has difficulty detecting inconsistencies in lengthy texts. On the other hand, human instructors take about 10 to 30 minutes to perform evaluations, providing detailed and individualized analysis. However, they can be more critical and may be influenced by personal biases and fatigue, potentially resulting in lower grades. ChatGPT can serve as an evaluation assistant, offering immediate and meaningful feedback, thereby reducing the workload.

The study by Jukiewicz (2024) found a strong positive correlation between grades given by ChatGPT and human instructors, with an insignificant difference between the two. ChatGPT-generated grades were slightly lower than those given by human instructors, as ChatGPT appeared to be stricter regarding programming assignment standards and more adept at detecting code errors. Human instructors tended to give higher grades to work that wasn't perfect, provided the code functioned and met the task requirements. This study evaluated assignments in a Programming course within the cognitive science program, using Python programming tasks.

Lu et al. (2024) evaluated the differences in feedback provided by ChatGPT and human instructors for academic writing tasks in Chinese. They found moderate to good consistency between the scores given by human instructors and ChatGPT (ICC = 0.6 and ICC = 0.75, respectively). ChatGPT provided more general and extensive evaluations, while human instructors offered specific explanations and solutions. Human instructor feedback was more frequently implemented by students (80.2%) compared to ChatGPT feedback (59.9%). The integration of ChatGPT in evaluations promoted a deeper understanding and independent thinking in student revisions, significantly improving their academic writing.

The study by Tossell et al. (2024) indicates that ChatGPT did not simplify students' writing tasks but changed how they perceive and approach assignments given by instructors, thereby improving their learning. Initially, students viewed ChatGPT as a fraudulent tool requiring human supervision, technical competence, and calibrated trust. After using it, students recognized it as a valuable learning tool, perceiving it as more ethical and benevolent. Despite this, they showed low comfort in taking responsibility for tasks completed

with ChatGPT's assistance due to ethical concerns and a lack of confidence in the accuracy of its results. Students preferred to be evaluated by both ChatGPT and the instructor, rather than by ChatGPT alone.

The study by Wang et al. (2024) evaluated ChatGPT's ability to provide feedback on university students' arguments and found that ChatGPT demonstrated high accuracy (91.8%) in evaluating quantitative points such as claims, evidence, and refutations, although its recall rate was 63.2%. ChatGPT's accuracy decreased with longer arguments and was influenced by the use of discourse markers. It provides more extensive, rapid, and text-based feedback, relying on data, but struggles to deliver affective feedback that is effective for students. In contrast, feedback from human instructors is more focused and based on experience.

## DISCUSSION

This systematic review analyzed eight studies comparing feedback generated by ChatGPT with that provided by human professors in higher education. The findings show that AI-generated feedback offers distinctive advantages in clarity, comprehensiveness, and immediacy. ChatGPT provides structured and extensive comments that allow students to promptly identify areas for improvement (Guo & Wang, 2023; Ivanovic, 2023). Its ability to analyze and produce feedback within seconds substantially reduces turnaround time compared to human instructors, who may require several minutes to review a single essay (Jukiewicz, 2024). This immediacy promotes iterative learning cycles, enabling students to apply feedback quickly and observe their progress in real time. Such efficiency supports self-regulated learning, a cornerstone of higher education, by empowering students to monitor and adjust their performance independently (Escalante et al., 2023). Additionally, the consistency and objectivity of AI-based feedback minimize discrepancies often observed among human evaluators, particularly in large classes where instructors face heavy workloads (Messer et al., 2024). These aspects collectively position ChatGPT as a potentially transformative tool for formative assessment, enhancing accessibility and timeliness in feedback delivery.

However, these technological strengths coexist with clear pedagogical limitations. Despite providing detailed and objective analyses, ChatGPT lacks the emotional intelligence and empathic communication that are fundamental to effective human feedback. Several studies reported that AI-generated comments, although comprehensive, sometimes fail to acknowledge students' affective needs or learning struggles, producing an impersonal experience that may reduce motivation (AlGhamdi, 2024; Wang et al., 2024). For learners with lower language proficiency, excessive or poorly contextualized feedback can even increase anxiety and hinder comprehension (Guo & Wang, 2023). This suggests that feedback is not merely a cognitive tool but also a social and emotional exchange that shapes students' confidence, self-efficacy, and persistence. Without empathy or encouragement, feedback, no matter how precise, risks becoming mechanistic and detached from the human dimensions of learning.

Human professors, in contrast, provide feedback enriched by context, experience, and emotional resonance. Teachers often tailor their comments to students' backgrounds and personalities, fostering trust and engagement. Their affective and motivational phrasing, acknowledging effort, improvement, and potential helps learners internalize constructive criticism and transform it into self-growth (Lu et al., 2024; Wang et al., 2024). Furthermore, human feedback can interpret subtleties of expression, creativity, and intent that AI systems cannot yet capture reliably. These qualitative dimensions of pedagogy are essential to the development of higher-order skills such as critical thinking, reflection, and ethical reasoning. Nevertheless, human feedback is limited by time, subjectivity, and variability between evaluators, which may compromise consistency and scalability in large academic environments.

The complementary nature of AI and human feedback emerges as one of the most significant findings of this review. Students perceive ChatGPT not as a substitute for human professors but as a collaborative tool that can reinforce and extend traditional feedback (Tossell et al., 2024). When used under professor supervision, AI can handle repetitive or technical aspects, such as grammar, coherence, or structure–while professors focus on higher-order elements like argument quality, originality, and conceptual depth (Escalante et al., 2023; Ivanovic, 2023). This division of cognitive labor aligns with contemporary educational theories that advocate human-machine symbiosis, where technology enhances rather than replaces pedagogy. Integrating both forms of feedback can increase fairness and timeliness while preserving the affective and contextual richness of professor-student relationships. However, such integration requires thoughtful pedagogical

design to prevent overreliance on automated systems and ensure that human judgment remains central in evaluating complex and creative work.

In practical terms, implementing a hybrid feedback model requires defining clear pedagogical roles for each agent. AI systems such as ChatGPT could be primarily employed for formative feedback, offering immediate, grammar or structure, related suggestions that support iterative learning. The human professors would remain responsible for summative evaluations and for addressing interpretative, ethical, and affective dimensions of student work. This division not only preserves academic integrity and emotional depth but also leverages AI's scalability to alleviate workload pressures in large classes. To ensure effective integration, institutions should promote structured frameworks that combine automation with human supervision, establishing guidelines for transparency, verification, and student agency.

The implications of these findings extend beyond classroom practice to institutional and ethical domains. Universities must establish clear policies regulating the use of AI tools, ensuring transparency, academic integrity, and data protection (Chan & Hu, 2023). Educators need training not only in how to use ChatGPT effectively but also in how to critically evaluate its output, recognizing potential biases and limitations in language generation. Ethical literacy in AI should become an integral component of teacher education programs, fostering awareness of issues such as fairness, accountability, and human oversight. At the same time, the academic community should resist the tendency to delegate all evaluative functions to algorithms, as this could diminish students' intellectual autonomy and critical reasoning abilities. Feedback should remain a dialogic process an exchange of meaning and reflection, rather than a one-way transmission of corrective information.

Future research should adopt longitudinal and mixed-method approaches to examine the long-term cognitive, emotional, and behavioral impacts of AI-assisted feedback on student learning outcomes. Understanding how sustained exposure to AI feedback influences academic identity, motivation, and metacognitive awareness is essential for developing robust and sustainable pedagogical frameworks. Comparative studies across disciplines are also needed to investigate how feedback practices vary between technical and humanistic contexts. From a broader pedagogical perspective, the integration of AI tools such as ChatGPT challenges educators to redefine the very purpose of feedback, shifting it from evaluation to dialogue, and from correction to student autonomy.

From a broader pedagogical perspective, the integration of AI tools such as ChatGPT challenges educators to redefine the very purpose of feedback, shifting it from evaluation to dialogue, and from correction to student autonomy.

It is also important to consider that demographic and contextual variables, such as academic discipline, cultural background, and prior exposure to AI tools, may influence how students perceive and respond to different feedback sources. Future studies should explore these factors to better understand cross-disciplinary and cross-cultural differences in the reception and effectiveness of AI-assisted feedback.

This systematic review has several limitations that must be considered. The included studies exhibit heterogeneity in design, sample size, and educational context, which may affect the generalizability of the findings. The rapid development of AI technologies such as ChatGPT also poses a methodological challenge, as research outcomes may quickly become outdated. Moreover, the emotional and contextual nuances that characterize human feedback are inherently difficult to quantify, complicating comparisons with AI-generated responses. Although most of the included studies were cross-sectional, some adopted short-term longitudinal or quasi-longitudinal designs, assessing feedback effectiveness or student perceptions over several weeks. Future research should therefore employ long-term longitudinal and mixed-method approaches to better integrate temporal, contextual, and affective dimensions of AI-generated feedback and its influence on learning outcomes.

Additionally, AI-generated feedback can occasionally produce inaccurate or contextually inappropriate suggestions, which may mislead students or distort their understanding of concepts. Its responses can vary across iterations, raising concerns about consistency and reproducibility. These challenges highlight the need for continuous human oversight and validation mechanisms when integrating AI into academic feedback systems.

Beyond methodological aspects, it is also crucial to examine the ethical implications and potential biases inherent in AI tools, ensuring that their implementation complements, rather than replaces, the human elements essential to education. Overreliance on automated systems risks diminishing the irreplaceable pedagogical and mentoring roles of human professors, potentially leading to depersonalized learning environments. In addition, the use of AI for summative assessments raises questions of fairness, transparency, and accountability, as algorithms may misinterpret student intent or context. Data privacy and informed consent also represent major ethical concerns, given that AI models often rely on extensive datasets that could include sensitive educational information.

Therefore, institutional policies must ensure data protection, academic integrity, and human oversight at every stage of AI use. While these technologies can enhance efficiency and accessibility, maintaining students' intellectual autonomy and preventing dependency on automated systems remain critical goals. Recognizing and managing the ethical challenges associated with AI use in education is fundamental. Thus, the findings indicate that technologies like ChatGPT should be regarded as complements, not replacements, to the human dimensions of education. A balanced combination of AI-driven efficiency and human empathy can cultivate more personalized, reflective, and ethically grounded learning environments in higher education.

## CONCLUSION

ChatGPT provides effective and timely feedback, enhancing learning through detailed and rapid responses. However, its lack of emotional nuance and specific guidance suggests it cannot completely replace human feedback in higher education. Integrating feedback from ChatGPT with that from human professors can optimize the learning experience. A hybrid approach that combines both forms of feedback may be the most effective strategy for improving educational outcomes in higher education. Students value the speed and detail of ChatGPT's feedback but prefer the personalization and empathy of feedback from human professors.

From a practical standpoint, ChatGPT can be integrated as a formative tool to support professors in large classes, provide immediate feedback, and promote student self-reflection. Nevertheless, its use should follow clear pedagogical objectives and ethical guidelines to maintain human-centered education. Future studies should investigate the long-term effects of AI-assisted feedback, its impact on students' motivation and autonomy, and its applicability across different disciplines and learning contexts. Combining AI efficiency with human empathy can create more inclusive, responsive, and ethically grounded learning environments in higher education.

**Author contributions: CHG-P:** conceptualization, writing – review & editing; **TC-L:** formal analysis, writing – original draft; **FMC-C:** methodology, writing – original draft, writing – review & editing; **RRI-D:** data curation, writing – original draft; **DVA-P:** data curation, writing – original draft; **NR-R:** data curation, writing – original draft; **BM-M:** conceptualization, writing – review & editing; **ÁT-P:** methodology, writing – review & editing; **RYFY:** data curation, writing – review & editing; **MP-M:** methodology, writing – review & editing. All authors approved the final version of the article.

**Funding:** This article was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES)–"Bolsista CAPES/BRASIL" finance code number 88887.967267/2024-00.

**Ethics declaration:** This study is a systematic review and did not involve the collection of primary data from human participants. Therefore, ethical approval from a Research Ethics Committee and informed consent from participants were not required.

**AI statement:** During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to assist with reviewing the clarity and readability of the text. All content was critically reviewed and edited by the authors, who take full responsibility for the final version of the manuscript.

**Declaration of interest:** The authors declared no competing interest.

**Data availability:** Data generated or analyzed during this study are available from the authors on request.

## REFERENCES

Al-Bashir, M., Kabir, M., & Rahman, I. (2016). The value and effectiveness of feedback in improving students' learning and professionalizing teaching in higher education. *Journal of Education and Practice, 7*(16), 38-41.

AlGhamdi, R. (2024). Exploring the impact of ChatGPT-generated feedback on technical writing skills of computing students: A blinded study. *Education and Information Technologies, 29*, 18901-18929. https://doi.org/10.1007/s10639-024-12594-2

Bauer, E., Greisel, M., Kuznetsov, I., Berndt, M., Kollar, I., Dresel, M., Fischer, M. R., & Fischer, F. (2023). Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. *British Journal of Educational Technology, 54*(5), 1222-1245. https://doi.org/10.1111/bjet.13336

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*, Article 43. https://doi.org/10.1186/s41239-023-00411-8

Cowling, M., Crawford, J., Allen, K. A., & Wehmeyer, M. (2023). Using leadership to leverage ChatGPT and artificial intelligence for undergraduate and postgraduate research supervision. *Australasian Journal of Educational Technology, 39*(4), 89-103. https://doi.org/10.14742/ajet.8598

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023a). Can large language models provide feedback to students? A case study on ChatGPT. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies* (pp. 323-325). IEEE. https://doi.org/10.1109/ICALT58122.2023.00100

Dai, Y., Lai, S., Lim, C. P., & Liu, A. (2023b). ChatGPT and its impact on research supervision: Insights from Australian postgraduate research students. *Australasian Journal of Educational Technology, 39*(4), 74-88. https://doi.org/10.14742/ajet.8843

ElSayary, A. (2023). An investigation of teachers' perceptions of using ChatGPT as a supporting tool for teaching and learning in the digital era. *Journal of Computer Assisted Learning, 40*(3), 931-945. https://doi.org/10.1111/jcal.12926

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*, Article 57. https://doi.org/10.1186/s41239-023-00425-2

Guo, K., & Wang, D. (2023). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*, 8435-8463. https://doi.org/10.1007/s10639-023-12146-0

Haindl, P., & Weinberger, G. (2024). Students' experiences of using ChatGPT in an undergraduate programming course. *IEEE Access, 12*, 43519-43529. https://doi.org/10.1109/ACCESS.2024.3380909

Hammoda, B. (2024). ChatGPT for founding teams: An entrepreneurial pedagogical innovation. *International Journal of Technology in Education, 7*(1), 154-173. https://doi.org/10.46328/ijte.530

Ho, W., & Lee, D. (2023). Enhancing engineering education in the Roblox metaverse: Utilizing ChatGPT for game development for electrical machine course. *International Journal on Advanced Science, Engineering & Information Technology, 13*(3), 1052-1058. https://doi.org/10.18517/ijaseit.13.3.18458

Ivanovic, I. (2023). Can AI-assisted essay assessment support teachers? A cross-sectional mixed-methods research conducted at the University of Montenegro. *Annales-Analiza Istrske in Mediteranske Studije-Series Historia et Sociologia, 33*(3), 571-590.

Jukiewicz, M. (2024). The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativity, 52*, Article 101522. https://doi.org/10.1016/j.tsc.2024.101522

Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education, 49*(5), 616-633. https://doi.org/10.1080/02602938.2024.2301722

Messer, M., Brown, N. C., Kölling, M., & Shi, M. (2024). Automated grading and feedback tools for programming education: A systematic review. *ACM Transactions on Computing Education, 1*(1), Article 1. https://doi.org/10.1145/3636515

Moola, S., Munn, Z., Tufanaru, C., Aromataris, E., Sears, K., Sfetcu, R., Currie, M., Qureshi, R., Mattis, P., Lisy, K., & Mu, P.-F. (2020). Chapter 7: Systematic reviews of etiology and risk. In E. Aromataris, & Z. Munn (Eds.), *JBI manual for evidence synthesis*. JBI.

Muniandy, J., & Selvanathan, M. (2024). ChatGPT, a partnering tool to improve ESL learners' speaking skills: Case study in a public university, Malaysia. *Teaching Public Administration, 43*(1), 4-20. https://doi.org/10.1177/01447394241230152

Ngo, T. T. A. (2023). The perception by university students of the use of ChatGPT in education. *International Journal of Emerging Technologies in Learning, 18*(17), 4-19. https://doi.org/10.3991/ijet.v18i17.39019

Ossa, C., & Willatt, C. (2023). Providing academic writing feedback assisted by generative artificial intelligence in initial teacher education contexts. *European Journal of Education and Psychology, 16*(2), 1-16. https://doi.org/10.32457/ejep.v16i2.2412

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Tong, S., Jia, N., Luo, X., & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal, 42*(9), 162-171. https://doi.org/10.1002/smj.3322

Tossell, C. C., Tenhundfeld, N. L., Momen, A., Cooley, K., & de Visser, E. J. (2024). Student perceptions of ChatGPT use in a college essay assignment: Implications for learning, grading, and trust in artificial intelligence. *IEEE Transactions on Learning Technologies, 17*, 1069-1081. https://doi.org/10.1109/TLT.2024.3355015

Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity, 51*, Article 101440. https://doi.org/10.1016/j.tsc.2023.101440

Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages, 8*(3), Article 212. https://doi.org/10.3390/languages8030212

Xu, X., Wang, X., Zhang, Y., & Zheng, R. (2024). Applying ChatGPT to tackle the side effects of personal learning environments from learner and learning perspective: An interview of experts in higher education. *PLoS ONE, 19*(1), Article e0295646. https://doi.org/10.1371/journal.pone.0295646

Yan, D. (2024). Feedback seeking abilities of L2 writers using ChatGPT: A mixed method multiple case study. *Kybernetes, 54*(7), 3757-3781. https://doi.org/10.1108/K-09-2023-1933

Yang, W., Lee, H., Wu, R., Zhang, R., & Pan, Y. (2023). Using an artificial-intelligence-generated program for positive efficiency in filmmaking education: Insights from experts and students. *Electronics, 12*(23), Article 4813. https://doi.org/10.3390/electronics12234813

Zhai, X., Chu, X., Wang, M., Tsai, C. C., Liang, J. C., & Spector, J. M. (2024). A systematic review of stimulated recall (SR) in educational research from 2012 to 2022. *Humanities and Social Sciences Communications, 11*, Article 489.

# APPENDIX A

**Table A1.** Search strategies applied in different databases (last search was carried out on 6 June 2024)

| Database | Search strategy | N |
|---|---|---|
| Scopus | TITLE-ABS-KEY ((“ChatGPT”) AND (“higher education” OR university OR “undergraduate students” OR college OR “Post-secondary education”) AND (“student feedback” OR “feedback” OR “AI-generated feedback” OR “Feedback sources” OR “Peer feedback”)) AND (LIMIT-TO (DOCTYPE , “ar”) OR LIMIT-TO (DOCTYPE, “re”)) | 45 |
| Web of Science | (“ChatGPT”) (Topic) and (“higher education” OR university OR “undergraduate students” OR college OR “graduate students” OR “Post-secondary education” OR education) (Topic) and (“student feedback” OR “feedback” OR “AI-generated feedback” OR “Feedback sources” OR “Peer feedback”) (Topic) and Article or Early Access or Proceeding Paper or Review Article (Document Types) | 92 |
| EBSCO | (“ChatGPT”) AND (“higher education” OR university OR “undergraduate students” OR college OR “Post-secondary education”) AND (“student feedback” OR “feedback” OR “AI-generated feedback” OR “Feedback sources” OR “Peer feedback”) | 31 |
| ACM Digital Library | [All: “chatgpt”] AND [[All: “higher education”] OR [All: university] OR [All: “undergraduate students”] OR [All: college] OR [All: “post-secondary education”]] AND [[All: “student feedback”] OR [All: “feedback”] OR [All: “ai-generated feedback”] OR [All: “feedback sources”] OR [All: “peer feedback”]]<br>Filters: Journals, Research article | 87 |
| IEEE Xplore | (“ChatGPT”) AND (“higher education” OR university OR “undergraduate students” OR college OR “Post-secondary education”) AND (“student feedback” OR “feedback” OR “AI-generated feedback” OR “Feedback sources” OR “Peer feedback”)<br>Filters: Journals, Early Access Articles | 11 |

## APPENDIX B

**Table B1.** Geographic region and sample size of the included studies

| Authors and year | Geographic region | Sample |
|---|---|---|
| AlGhamdi (2024) | Saudi Arabia | 111 first-year computing students, all men, aged 19 to 21. |
| Escalante et al. (2023) | USA | 43 undergraduate students in the Asia-Pacific region, registered in an English course, B1 proficiency level (CEFR). 13 M, 30 F, aged 19-36. |
| Guo and Wang (2023) | China | 50 undergraduate students, registered in an English course, with B2-C1 proficiency level (CEFR), 24 M, 26 F, aged 18-21, mean 19.54 years |
| Ivanovic (2023) | Montenegro | 78 students at the faculty of science and mathematics, 34 M, 44 F |
| Jukiewicz (2024) | Poland | 67 students. 25 students completed 9 projects. 20 completed 8 projects, 16 completed 7 projects. |
| Lu et al. (2024) | China | 46 education students of academic writing training program, 4 M, 42 F, 23.35 years. |
| Tossell et al. (2024) | USA | 24 senior-year students from an engineering course at the United States Air Force Academy with limited experience with ChatGPT. 16 M, 8 F, 22.25 years. |
| Wang et al. (2024) | China | 42 second year students in an argumentation teaching activity |

Note. CEFR: Common European framework of reference for languages; F: Female; M: Male

# APPENDIX C

**Table C1.** Characteristics of the included studies

| Authors & year | Type of work evaluated | Exhibition & control details | Data collection time | Variables evaluated | Results |
|---|---|---|---|---|---|
| AlGhamdi (2024) | Essay submitted on the "Blackboard weekly writing assignments" blog as an integral part of the course curriculum. | Human feedback and feedback generated by ChatGPT | 6 weeks. 1st-3rd weeks (professor feedback) 4th and 5th weeks (ChatGPT-generated feedback) | Emotional and psychological responses. Perceived quality and usefulness. Progress and development. Content and delivery. | 109 students (1st-3rd weeks), 102 students (4th-5th weeks), reactions vary, from the need for empathy and clarity in feedback. Feedback helps to improve and learn. Others criticize the lack of customization and consistency. It must be precise and empathetic, considering the emotional and psychological spaces of the students. |
| Escalante et al. (2023) | Writing of texts in English | Human tutors and ChatGPT-4 | 6 weeks. Weekly assessment | 8 items, feedback, divided into 4 groups: satisfaction, clarity, usefulness, preference: 1 to 5. | 18 students preferred to receive feedback from human tutors (mainly in the items of satisfaction, clarity, kindness, and preference). 20 students preferred AI. There was no significant difference. 5 students did not respond to the final survey. |
| Guo and Wang (2023) | 300-word English text composition after class | 5 human instructors (10 essays for each) and ChatGPT | 1 class | Content: quality and development of arguments. Organization. Language: spelling, format | Professors paid more attention to content and language and less attention to organization. More directive and informative comments. ChatGPT provided a greater number of comments across all three evaluated parameters. More directive comments and praise. |
| Ivanovic (2023) | English text composition | 3 human professors and ChatGPT 3.5 | Students had 2 months (December 2022-January 2023) | Quality of its content. Strength of arguments. The use of evidence. Relevance of the content. | Calculated ICC value of 0.8 between professors and ChatGPT indicates good consistency and reliability among evaluators. |
| Jukiewicz (2024) | The course "programming," in the cognitive science program. The number of tasks for assessment was 1,579. | Human instructor and ChatGPT 3.5-turbo (each evaluated 15 tasks). | 15 weeks (2×/week, 1.5 hours) | ChatGPT prompt engineering for developers was used. The work is evaluated as correct, almost correct, or incorrect: 1, 0.5, or 0 points. | In all tasks, the average scores from the professor are slightly higher than those from ChatGPT. The standard deviation of the professor's scores is greater, indicating more variation in the ratings. ICC of the 15 responses from ChatGPT is 0.13, indicating an insignificant difference between the responses. |
| Lu et al. (2024) | A 300-word summary of a fictional article about contemporary challenges in Chinese language teaching classrooms | Human professors and ChatGPT 3.5 | 6 weeks (1×/week, 3 hours). | -Ability to organize content logically: five levels (0-8)<br>-Express content concisely: five levels (0-8). Max. total score: 40. | Moderate to good coherence between the scores of the professor and ChatGPT (ICC 0.6 to 0.75). ChatGPT provided extensive comments and general suggestions. The professors included more praise, explanations, and specific solutions. |
| Tossell et al. (2024) | Writing assignment on the current challenges of human factors and human-computer interaction. | Human instructors and ChatGPT 4.0 | Approximately 2 months | - Perceived quality and difficulty: 1-7<br>- Educational value and level of comfort being responsible for the ChatGPT text: 1-7<br>- Perceived reliability: 0-7<br>- Confidence in text: 0-7<br>- Confidence in the evaluation: 1-7<br>- Evaluation preference: ChatGPT, instructor, or both | - Quality: before 5.48, after 4.75<br>- Difficulty: before 4.8, after 5.25<br>- Educational value: before 5.43, after 5.57<br>- Responsibility: before 3.71, after 3.82<br>- Reliability: difference between pre and post of the ethical and benevolent subscale<br>- Confidence in text: before 4.07, after 4.23<br>- Confidence in evaluation: instructor 6.29, ChatGPT 4.29, both 5.5<br>- Evaluation preference: 15/24 preferred instructor, 9 preferred instructor and ChatGPT |
| Wang et al. (2024) | 50 argumentation contents that human teachers had previously assessed. 84,000 words. Short arguments: 13 (800-1,300 words). Medium arguments: 16 (1,300-1,800 words). Long arguments: 21 (1,800-2,300 words). | Human professors and ChatGPT 3.5 | 8 weeks | Evaluation dimensions: claim, evidence, rebuttal, adequacy of evidence and explanation. | -Precision rate 91.8%. Claim (100 %), evidence (95.8 %), and rebuttal (91.0 %), the adequacy of evidence (85.3 %), explanation (85%).<br>-Recall rate 63.2%. Claim (100 %), evidence (89.2%), rebuttal (75.9 %), the adequacy of evidence (47.4%), explanation (29.8 %) |

# APPENDIX D

**Table D1.** Comparative analysis of professor and ChatGPT feedback

| Authors & year | Feedback | Time efficiency | Detail | Specificity | Clarity | Emotional & psychological implications |
|---|---|---|---|---|---|---|
| AlGhamdi (2024) | Professor | Time efficiency is limited | Variable, sometimes limited, with brief comments and general guidance | General and guiding, focused on general recommendations regarding organization, style, and clarity | It may be ambiguous | It inspires students to strive for excellence. It increases confidence and enthusiasm to continue writing |
| | ChatGPT | High time efficiency | Elevated, systematic, structured in organized lists. | High difficulty allows the student to identify the exact sections containing the error in order to correct it | Clear and well structured, facilitating the understanding of what needs to be modified and the underlying rationale | It stimulates self-reflection and growth. It provokes irritability when noticing numerous errors in the writing |
| Escalante et al. (2023) | Professor | It requires a substantial amount of time | Adequate, but dependent on time availability and synchronous interaction. It focuses on the most relevant points of the text | It is contextualized and interpretative, but may be less specific in linguistic or microstructural terms | Generally clear, especially when face-to-face interaction is available, allowing immediate clarification and reformulation | Relevant affective benefits, such as engagement, motivation, a sense of support, and opportunities for dialogue |
| | ChatGPT | It shows high time efficiency | Highly detailed, including systematic comments by category, accompanied by examples | High specificity, identifying errors sentence by sentence, error types, metalinguistic explanations, and suggested revisions | Very clear, well organized, and easy to understand | It lacks human interaction and emotional sensitivity, with a risk of losing the text's personal voice and a weaker affective bond |
| Guo and Wang (2023) | Professor | 48 to 305 minutes for 10 essays | Selective and focused feedback, prioritizing content and language aspects | Informative and interrogative feedback, fostering student reflection and autonomy | Dependent on individual teaching style and experience, and may be enhanced through oral interaction or in-class discussion | It provides emotional support, empathy, and adaptation to the student's profile, fostering motivation and acceptance of the feedback |
| | ChatGPT | A large volume of feedback delivered within a few seconds | More extensive, systematic, and detailed feedback, with justifications for the suggested revisions | Directive feedback, indicating exactly what should be modified, often with concrete examples | Generally clear, structured, and logical, with explicit clarification of the purpose of the suggestions | It may include initial praise; however, excessive feedback and a lack of sensitivity can lead to anxiety, cognitive overload, and confusion |
| Ivanovic (2023) | Professor | 5-15 minutes | Feedback varies in its level of detail, often constrained by the available time and the repetitive nature of the task | Interpretative and global comments, with less emphasis on microstructural aspects in some cases | Clarity dependent on cognitive state and fatigue; it may vary throughout serial grading | Capable of recognizing students' effort, difficulties, & emotional state, adjusting the tone according to the student's profile |
| | ChatGPT | < 30 seconds | Systematic and detailed feedback, covering grammar, coherence, structure, and arguments in a standardized manner | High specificity, with clear identification of weaknesses, examples of problems, and direct suggestions for improvement | High clarity and consistency, with explicit language and stable logic across different texts and evaluated submissions | It lacks empathy, emotional intelligence, and human pedagogical sensitivity |
| Jukiewicz (2024) | Professor | Several hours to correct dozens of tasks | Evaluation focused on overall functioning of the code, without identifying subtle flaws or specific requirement omissions | It may be influenced by subjective judgment and tolerance of partially correct solutions | Feedback may be brief or poorly structured, depending on the available time, and may vary significantly across students or tasks | Capable of exercising pedagogical empathy, considering the student's effort and educational context |
| | ChatGPT | 9.5 seconds per task | High capacity for detailed analysis, identifying logical errors, requirement omissions, and code quality aspects | It applies explicit and consistent criteria, comparing the student's solution with an ideal solution generated by the model itself | Clear, structured, & standardized feedback, explaining rationale for assigned grade and suggesting specific improvements to code. | Emotionally neutral assessment ensures objectivity; but it eliminates empathic and relational dimension of the educational process |

**Table D1 (Continued).**

| Authors & year | Feedback | Time efficiency | Detail | Specificity | Clarity | Emotional & psychological implications |
|---|---|---|---|---|---|---|
| Lu et al. (2024) | Professor | They require a high investment of time | More detailed feedback on critical aspects of the text, with greater emphasis on explanations, specific solutions, and clear justifications for the assigned grades | Greater specificity, providing clear and targeted instructions for revision, which facilitates comprehension | Clearer, more legible, and easier to understand, promoting greater agreement with and acceptance of the guidance provided | It includes a greater number of compliments and motivational comments, strengthening rapport, confidence, and emotional engagement |
| | ChatGPT | Highly time-efficient | It produces a greater overall quantity of feedback, with general comments and suggestions, without delving into specific issues | Lower specificity, with suggestions that are often abstract or vague | Longer and, in some cases, difficult to understand or apply | Few compliments and limited encouragement. The absence of human interaction restricts affective support, although it may stimulate a more critical and reflective stance |
| Wang et al. (2024) | Professor | They require considerable time. The process is slower | Deeper and more qualitative feedback, addressing argumentation, coherence, originality, and alignment with the course content | It provides more specific and targeted comments, facilitating feedback implementation and text revision | Clearer and more reliable. Students better understand expectations and assessment criteria | It plays a central role in emotional and motivational support. Human presence enhances the sense of fairness, trust, and legitimacy of the assessment process |
| | ChatGPT | It provides near-immediate feedback, being able to analyze texts and generate comments within seconds | It produces a large volume of comments but often delivers repetitive or superficial responses | It tends to provide generic or abstract suggestions. The lack of precision hinders direct application of the feedback | Although linguistically clear, the feedback may raise concerns regarding its reliability and accuracy | Emotionally neutral. Although perceived as ethical and benevolent, it does not replace the psychological support and sense of security provided by the instructor, potentially causing discomfort when used in isolation |
| Tossell et al. (2024) | Professor | It requires a high time investment (≈15-20 minutes per assignment) | More focused and selective feedback, directed at the most critical issues in the student's argumentation, based on pedagogical experience and academic history | Greater contextual specificity, identifying the student's immediate and recurring problems, with guidance directly applicable to performance improvement | Generally concise, clear, and easily understandable, facilitating accurate interpretation of guidance and greater acceptance by students | A strong humanistic and empathic component, with praise, encouragement, and consideration of the student's progress, contributing to motivation, confidence, and self-efficacy |
| | ChatGPT | Instant feedback | Comprehensive and detailed feedback, covering multiple assessment points (claims, evidence, rebuttals), including aspects that the instructor may not mention | High specificity in quantitative evaluation (identification of claims, evidence, and rebuttals), but lower precision in qualitative assessments (adequacy of evidence and explanations) | Linguistically clear and well-structured feedback; however, it is extensive. In long or complex texts, it may present omissions or a loss of conceptual clarity | Affective feedback limited to the linguistic level (standardized praise such as "good job"), without genuine empathy or understanding of the student's emotional and historical context |

# APPENDIX E

**Table E1.** Advantages and disadvantages of professor and ChatGPT feedback

| Authors & year | Feedback | Advantages | Disadvantages |
|---|---|---|---|
| AlGhamdi (2024) | Professor | Contextualized and individualized feedback, delivered with pedagogical and emotional sensitivity, fostering engagement, motivation, and a clear understanding of academic expectations | Feedback may be inconsistent, constrained by available time and class size, resulting in reduced detail and specificity. Variability in assessment |
| | ChatGPT | Precise and detailed feedback. A corrective tool that standardizes evaluative criteria and reduces faculty workload | Overly standardized feedback, with limited personalization, and lacking empathy toward students' emotional states |
| Escalante et al. (2023) | Professor | Human interaction, the possibility of immediate questions, emotional support, and pedagogical contextualization. | High time cost, limited scalability, and potential variability among evaluators |
| | ChatGPT | Clear, consistent, detailed, and specific feedback, with potential for standardization and scalability | Absence of dialogic interaction, limited emotional personalization, dependence on prompt quality, and excessive standardization |
| Guo and Wang (2023) | Professor | Knowledge of the pedagogical context, curriculum, and students' academic history, along with the capacity for personalization, dialogue, and emotional mediation | Time constraints, lower standardization, and difficulty in addressing all aspects of writing in large classes |
| | ChatGPT | Balance among different dimensions of writing, clarity, explicit justification for suggestions, and potential to reduce faculty workload | Excessive feedback, occasionally irrelevant, and difficulty in locating issues in the text due to a lack of contextual knowledge and course-specific pedagogical criteria |
| Ivanovic (2023) | Professor | Capacity for pedagogical contextualization, human critical judgment, cultural sensitivity, and interpretation of creative nuances | High time cost, susceptibility to mental fatigue, and intra- and inter-rater variability |
| | ChatGPT | Significant reduction in assessment time, immediate feedback, standardization, and direct support for the instructor | Absence of empathy, limited understanding of creativity and sociocultural context, and potential biases inherent to the training data |
| Jukiewicz (2024) | Professor | Didactic experience, contextualized human judgment, adaptability, and a comprehensive understanding of the student's learning process | Subject to subjectivity, inter-rater variability, fatigue, and occasional grading errors, especially in extensive and repetitive assessments |
| | ChatGPT | High time efficiency, strong correlation with instructors' grades, rigorous application of programming standards, and automatic generation of detailed feedback | Financial cost, potential hallucinations, and the need for multiple interactions or instructor supervision to ensure fair assessment |
| Lu et al. (2024) | Professor | High credibility, clarity, specificity, and a higher rate of feedback implementation | A high workload and time constraints may compromise the depth and frequency of individualized feedback |
| | ChatGPT | Good consistency with instructor assessment, a greater volume of feedback, and encouragement of independent thinking, criteria comparison, and the development of feedback literacy | Lower feedback implementation rate, greater student rejection, risk of imprecise or abstract comments, and absence of emotional interaction |
| Wang et al. (2024) | Professor | Greater student trust, clearer assessment criteria, higher feedback specificity, and greater legitimacy in assigning final grades | A high workload and time constraints may limit the frequency and extent of individualized feedback |
| | ChatGPT | Useful as a collaborative learning tool. It promotes critical thinking, active revision, and AI literacy | Low confidence for standalone assessment, risk of imprecision, repetitiveness, lack of specificity, and absence of human interaction |
| Tossell et al. (2024) | Professor | Pedagogical experience, empathy, cumulative formative assessment, and the ability to integrate cognitive and emotional aspects of the learning process | Time limitations, difficulty in providing immediate feedback, and risk of subjectivity or inconsistency when handling large volumes of assignments |
| | ChatGPT | High overall accuracy, excellent performance in the quantitative evaluation of arguments, and immediate, comprehensive, and scalable feedback | Lower recall, reduced performance in qualitative assessments and long texts, dependence on discourse markers, absence of human empathy, and risk of incomplete or superficial feedback |