ChatGPT-Generated Reviews for University Students' Papers – How are they Perceived?

Daniel Staegemann
Otto-von-Guericke University Magdeburg
Magdeburg, Germany *daniel.staegemann@ovgu.de*

Klaus Turowski
Otto-von-Guericke University Magdeburg
Magdeburg, Germany *klaus.turowski@ovgu.de*

Abstract

With research proving the value of intensive tutoring for learners to improve their results but university educators often being limited in time and numbers, harnessing large language models (LLM) to bridge this gap and provide more support appears to be an auspicious solution. One promising avenue for this is students' training in writing, since producing coherent texts is a strength of LLMs. To explore how university students and teachers perceive the quality of ChatGPT-generated reviews, a study was conducted in the frame of a university course on scientific writing for IT students, identifying the potential benefits but also the weaknesses of using ChatGPT as a reviewer for university students' scientific texts.

Keywords: Large language model, LLM, higher education, experiment, perception.

1. Introduction

Since ChatGPT was presented to the public in 2022 [19], large language models (LLM) have gained vast public interest. This also extends, but is not limited to, academia, where numerous avenues to harness their capabilities are explored [6], [20]. Hereby, their capability to produce complex texts based on provided inputs positions them as a promising tool for boosting productivity in numerous domains [5], [8], [24].

One of the fields where their utility is currently extensively explored is the educational sector. Here, particularly the scarcity of teaching personnel and their lack of time to individually and thoroughly tutor each individual learner, which many higher education institutions are affected by, especially those with fewer financial resources, could potentially be addressed by LLM-based approaches. While this issue is relevant across subjects and tasks, since individual support is helpful in most scenarios [3], it particularly applies to highly individualized student-tasks with a high degree of complexity. Thus, learning to write (high-quality) texts is a prime example of a learning-activity that could highly benefit from virtual tutors. Not only are the texts that shall be produced individual, causing an increased effort for meaningfully commenting on them, but high-quality writing is also often an iterative process [15], which leads to the necessity of assessing the produced text multiple times, until the end result is satisfactorily. This appears even more relevant for students that oftentimes do not get much writing experience as part of their curricula, such as, for instance, in the IT domain. While university instructors may lack the time to give meaningful feedback for each iteration and also only have limited working hours [4], [12], using a LLM instead could alleviate both issues. However, for this to be feasible, the quality of the provided commentary needs to be of sufficient quality, especially when compared to human-written reviews. Since ChatGPT is the most popular LLM with the highest user base [27], for this study it was also used as the LLM of choice. Therefore, the study that is presented in the publication at hand aimed to answer the following research question (RQ):

RQ: *How do university teachers and students perceive the quality of ChatGPT-generated reviews for IT students' scientific papers compared to reviews that are written by humans?*

To answer the RQ, the remainder of this paper is structured as follows. After this introduction, the background section provides a foundation for the experiment that is described subsequently. Afterwards, the findings and limitations are discussed and then a conclusion including potential avenues for future work is given.

2. Background

Due to their capabilities in producing naturally sounding texts based on provided inputs [30], the general idea of using LLMs as a reviewer is rather obvious. Thus, several works are already dedicated to exploring its feasibility, of which several will be presented in the following, to provide an impression of the domain, without claiming comprehensiveness.

How well LLMs perform at evaluating student project reports was explored in [7]. Here, similar to the study presented within this paper, despite the different context of the analyzed student works, the motivating factor was also the desire to provide timely feedback to students' submitted work. Yet, in the evaluation, the examples were considered individually and not ranked in direct comparison. The same applies to [28], where the authors compared different LLM-based approaches for grading argumentative essays and also tasked the LLMs to provide a reasoning for their scoring. In both cases, however, the focus was on the evaluation and not on providing guidance for improvement

The capability of ChatGPT to evaluate research papers was explored in [26]. However, there it was tasked to provide a rating and textual evaluation, but not recommendations for improvement. Something similar was also done in [29], where ChatGPT was also used to generate ratings and textual reviews. Yet, again, the latter was focused on the evaluation of the provided text's quality and not on providing suggestions for its improvement. These suggestions were an important factor in [16], where ChatGPT's feedback on English-as-a-foreign-language writing assessments was explored. Hereby, the quality of its comments was found to even surpass those by humans.

In another study design, in [17], researchers were asked to decide for pairs of texts, which one was written by ChatGPT and which was written by a human. Thus, in contrast to the study at hand, they were aware of the presence of ChatGPT-texts. Nevertheless, the researchers only decided correctly 50 percent of the time, highlighting ChatGPT's ability to produce human-sounding texts. A similar detection rate was also determined in [14], where professors from different disciplines were asked to assess for numerous texts, if these were written by students or by AI (in the given case, ChatGPT 3.5 was used).

Overall, it is evident that the general topic of the publication at hand is very present in the literature, highlighting its relevancy. However, the findings are somewhat mixed, showing potential but also considerable weaknesses in the explored settings. Since the widespread research on LLMs is just in its infancy, adding additional insights and examples to the body of literature appears reasonable and necessary. Considering specifically the focus of the publication at hand, such a paper with a focus on the perceived value of LLM-created reviews compared to their human-made counterparts for university students' papers in the IT domain was not found. Yet, students in this field generally don't have many writing-based assignments which limits their exposure and makes obtaining frequent and high-quality feedback and suggestions for improvement even more important. Thus, in the following, this topic will be explored to add to the overall understanding of how LLMs can be harnessed in (higher) education.

3. The Experiment

As highlighted earlier, the goal of this study was to explore university teachers' and students' perception of ChatGPT-generated reviews compared to those written by humans. For this purpose, an experiment was conducted that will be described in the following, outlining the general setup, the creation of the ChatGPT-reviews, and the obtained results.

3.1. The Setup

The conducted study took place in a German university's master-level course on academic writing for IT students (students from degrees such as computer science, business informatics, or data and knowledge engineering). More specifically, the course focused on the students writing scientific papers of their own on topics of business informatics or related domains. Due to the level of the course, all the students had several years of experience in university studies, already obtained a bachelor's degree and, thus, passed the associated bachelor's thesis, and in many cases, they already had noteworthy working experience. Hence, the participants can be considered generally somewhat experienced in the IT domain. However, based on their academic background, it can be expected that their writing-experience was rather limited compared to, for instance, students of the humanities. Further, the university teachers involved had between 1 and 20+ years of teaching experience (in several cases also in previous iterations of the course) and are also actively publishing and reviewing scientific papers themselves.

In the course that lasted one semester, eleven students were tasked to write a scientific paper (in contrast to, for instance, prose) on a topic provided by the teachers. Even though the papers are primarily intended as an academic exercise, to increase the students' motivation, it was clearly communicated that the aspiration was to refine and publish the best results in the aftermath of the course.

The papers' length was supposed to be twelve pages in a given template [11]. While the topics provided a direction, within their boundaries, the students could shape the concrete focus and pursued research questions. Due to the request of two students, one topic was based on the design science research [10] paradigm, and was given to them as a team, whereas the others all worked individually on papers that presented structured literature reviews. Hence, overall, ten papers were created. Every paper's focus was unique and different topics within the overarching theme were covered. The course's grade was determined based on three components, the final paper (65%), reviews written for other students' papers (20%), and a presentation of the research project (15%). This combination aimed to provide the students with experience in scientific writing as well as to introduce them to the activities that are associated with scientific publications.

In the first two weeks, the students were given lectures (two slots with a duration of three to four hours each) that outlined the principles of scientific work and different research methods. Further, in the first week, the available topics were introduced, and the students were asked to apply for the topics by stating their preferred order of topics and writing a brief (up to half a page) motivation. Then, in the second week, the topics were assigned based on the students' preferences and, in the case of several applicants for one topic, the merit of their application. Subsequently, over the majority of the semester, the students wrote their papers. For this, they were suggested milestones that should be reached at certain points. During this time, every second week, there was a session where students were asked to present their progress, followed by a discussion with the present teachers (always at least two) and other students. However, neither adherence to the milestones nor participating in the presentations, or the use of the offered consultations were enforced, giving the students a high degree of freedom. Towards the end of the semester, the students should submit their first completed draft for the purpose of getting reviews. While the quality of this draft did not impact their grade and, theoretically, the invitation for the submission could have been ignored, seven individual students and the team followed through, appreciating the value of obtaining the feedback, whereas two students waived the possibility. Thus, eight drafts were submitted.

Subsequently, each student was assigned three papers for review, for a total of 33 student reviews. As the students knew in advance that the quality of their review impacts their grade and it was also highlighted that their peers would heavily benefit from the input for their final version of the paper, it could be assumed that all students would take the task seriously. Further, the students were also invited to review additional papers if they wanted to gain more corresponding experience. While this was primarily done to further obfuscate the appearance of additional (ChatGPT) reviews, one student actually used the option, bringing the total number of student reviews to 34. In theory, the reviews were supposed to be double-blind. However, due to the status updates during the semester, the students most likely knew who authored which paper. Yet, this was no issue, since the draft version (and thus the assessment provided by the reviews) were not part of the grade. Besides the student reviews, also the teachers provided three reviews each. Within the course, besides this paper's authors, five further teachers were involved and an additional researcher also provided reviews, resulting in a total of 21 reviews by the instructors. Hereby, three of the reviews were written by this publication's first author, whereas the second author was overall responsible for the course but did not provide any reviews. For writing the reviews, the students were given one week.

The assignment of the reviews was determined by this paper's first author. The goal was to ensure a fair distribution regarding the content (complexity of the topic and reviewer's familiarity with it), the paper's expected quality (based on the impression of the presentation settings), and the expected review quality provided by the participants (based on the impression of the presentation settings and consultations as well as the status as student or teacher) as well as to keep people with known relations separated to increase objectivity. While this was naturally far from perfect, since, for instance, the predictions of paper and review quality were highly speculative, overall, every draft was provided with valuable feedback across the total set of reviews. Hereby, everyone received between three and five student reviews and two or three teacher reviews. However, besides this paper's authors, no one knew the assignment of the reviews and students and teachers were requested to neither discuss the assignment nor the reviews themselves. While this was publicly justified with scientific integrity and the fact that reviews in the real scientific world are also not to be discussed, the actual main purpose was to inconspicuously inject a review from ChatGPT. This is because to facilitate the presented study, without the knowledge of the teachers or students, each student paper was also provided with a review written by ChatGPT (more details on this are given later in the paper). Therefore, in total, each student (team) received seven or eight reviews. These were all consolidated into one PDF and sent to the respective draft-authors. For the teachers, all these files were made accessible in a shared folder. Importantly, to avoid influencing the verdict, neither the students nor the teachers were told that there is a ChatGPT-generated review. While it is usually customary to disclose details regarding an experiment to the participants, deviating from this practice is seen as acceptable if it is necessary for the experiment's success (as in the given case) and does not risk to cause harm to the participants as highlighted in points 8.05 and 8.07 of [2] and also practiced in other studies, especially in the context of human-computer interaction [9], [13], [18], [21, 22]. Thus, the participants were only informed and asked for agreement after the conclusion of the experiment.

Once all reviews were distributed to the students who authored the respective papers, they could improve their paper drafts before the final submission. For this, they had one week. While it would have been interesting to determine the impact of the reviews by ChatGPT on the final papers, this was not possible with the current setup (if it is at all) and was, therefore, also not attempted. Instead, the students were asked to provide their assessment of the reviews they got for their draft together with their final paper submission. To increase the likelihood of getting an honest assessment, they were made aware that the grade for the reviews would be solely determined by the teachers. Further, since they were only asked for a ranking and not for an absolute assessment, it was not possible to just rate everyone favorably. In theory, students could have communicated with each other to figure out which reviews were by students and which by teachers. However, looking at the received rankings, this apparently did not happen and would have also served no purpose, since it wouldn't have impacted the grades. Unfortunately, one student did not submit their ranking. Hence, only eight student review-rankings for seven papers were submitted. Hereby, the members of the one team were requested to create their rankings separately, however it can be assumed that communication took place, which might have influenced the individual assessments. Furthermore, the teachers that participated in the course were also asked to provide their review-rankings for all papers. Thus, for each paper, the reviews were ranked by five university teachers. Additionally, this publication's first author also created such rankings (attempting to also rank the ChatGPT-reviews as objective as possible), yet, due to the knowledge of the ChatGPT review and the involvement in its creation, these rankings have to be treated differently from the other ones and were, therefore, not included in the analysis.

After submitting their final papers and the rankings of the reviews, the students presented their research at a mock conference and, subsequently, received feedback and their grades.

However, this is not elaborated further, since it is not relevant for the purpose of this paper. Following the course, as depicted in Fig. 1, questionnaires were distributed to the participants, asking them about their perception of the ChatGPT-reviews and ChatGPT as a reviewer. While the findings do not allow for a comprehensive analysis, since the number of returned surveys was even lower than the number of course participants and the thoroughness in filling out the surveys was rather heterogeneous, they still provided some impression of the perception.

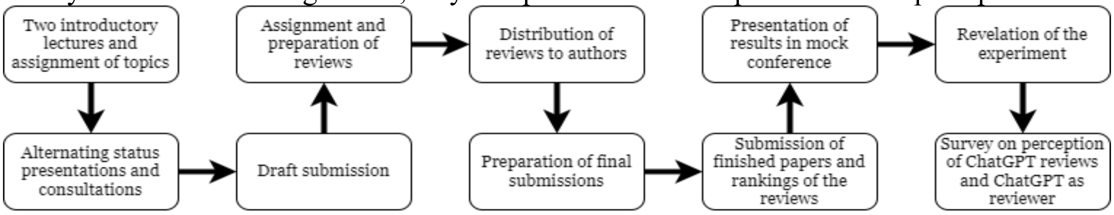
Two introductory lectures and assignment of topics
Alternating status presentations and consultations
Draft submission
Assignment and preparation of reviews
Distribution of reviews to authors
Preparation of final submissions
Submission of finished papers and rankings of the reviews
Presentation of results in mock conference
Revelation of the experiment
Survey on perception of ChatGPT reviews and ChatGPT as reviewer

Fig. 1. Sequence of the course and experiment.

3.2. Generation of the ChatGPT-Reviews

To generate the provided ChatGPT-reviews, ChatGPT-4o was used, promising a high quality in comparison to older versions. However, even though other strategies might generally improve the quality of ChatGPT-generated responses [1], [25], zero-shot prompting was utilized for several reasons. On the one hand, many reviewers have their own way of approaching the task and structuring their texts, which would have been hard to capture. On the other hand, this way, it could not only be explored how well ChatGPT fares in providing reviews in general but also how useful it would be to recommend students to use it for getting immediate feedback on their current drafts. While students could, of course, also utilize sophisticated prompting strategies, the likelihood that the majority resorts to rather simple prompts was seen as high, which this was also emulated for this experiment.

However, an important goal of the reviews was to facilitate the improvement instead of just providing a mere assessment of the quality. After all, the students were supposed to use the reviews to improve their draft in the course of preparing their final submission. Therefore, this focus was not only communicated to the students and teachers but also included in the prompt given to ChatGPT. Moreover, to indicate the expected level of sophistication of the review and to emulate the experience of the involved teaching staff in this course as well as the assumed experience of teachers in other courses on scientific writing, the prompt explicitly stated that the review should be suitable for a scientific conference.

Additionally, some instructions regarding the length and formatting were given to avoid too big differences from the expected human-written reviews. Further, since some students deleted the template's part regarding the authors and affiliation, others filled in placeholders. To ensure consistency and avoid confusion, ChatGPT was explicitly ordered to ignore these as well as the names of the provided files. Incorporating all the above considerations, the final prompt that was used to create reviews for all the provided student-drafts was as follows:

Please write a review for the uploaded paper, focusing primarily on potential ways to improve the paper rather than describing its content. The review should be suitable for submission to a scientific conference and should avoid sub-headings within the text. The review shouldn't mention the author's name or affiliation. The review should not mention the filename but it can mention the paper's title. The review should have at most five paragraphs. The review should have a length of 0.6-1.0 A4 pages.

While it was initially intended to pass the reviews to the students without making any changes, this plan had to be slightly altered. Even though the LLM was explicitly told to structure the text in five or less paragraphs, it ignored this instruction. Instead, the reviews were partitioned in many small abstracts, which appeared rather unnatural. Thus, this paper's first author reformatted all the reviews that were created by ChatGPT. However, this was the only intervention, and the content was not altered at all. Furthermore, the reviews by humans were not changed at all, therefore, there was still a plethora of styles in each set of reviews. Theoretically, the consistency in style across all ChatGPT-reviews could have been picked up on by the teachers who saw all reviews for all papers (whereas the students only saw the reviews for their draft and, thus, only one ChatGPT review), yet, this apparently did not happen as evidenced by the provided ratings as well as conversations that were held with them after revealing the experiment.

3.3. Results

As mentioned earlier, each student was asked to rank the reviews they received for their draft, while the involved teachers should provide separate rankings for all of the drafts‘ reviews. The submitted rankings are shown in Table 1, where each part of the table contains the results for one of the drafts. Hereby, the columns *T1* to *T5* show the assessment of the involved teachers, whereas *Student* shows the student's ranking. In the case of Table 1, two student ratings are given, since this was the team's paper and both submitted their individual rankings that are displayed here. The reviews by the students are depicted with the IDs *S1* to *S11*, depending on the author, with each student having a distinct ID. Further, the IDs are consistent throughout all of the tables, allowing to compare the placement of each student's reviews across all the papers they reviewed. The entry *C* denotes the reviews written by ChatGPT. Moreover, the teacher IDs are also the same for the column heads. Besides the five teachers that were fully involved, one additional instructor provided reviews, but no rankings and this paper's first author provided reviews but their assessments were left out, due to the knowledge about the experiment. Thus, *T6* and *T7* appear in the cells but not in the column heads. Unfortunately, the author of *Draft 8* did not submit their rankings, which is why the corresponding column contains no values. For better visual clarity, the cells are color-coded depending on the type of review (teacher, student, ChatGPT).

Table 1. Rankings of the reviews for the eight paper drafts

Rank	Draft #	T1	T2	T3	T4	T5	Student	Draft #	T1	T2	T3	T4	T5	Student
1	Draft 1	T5	T5	T6	T5	C	C |T4	Draft 2	T6	T1	T6	T1	T1	T6
2		C	C	T5	T6	T6	T4 | C		C	T5	T1	T5	C	T1
3		T6	T6	C	S7	S9	S6 |S6		T5	T6	C	S2	T6	T5
4		S4	S9	S4	S9	S4	T6 |T6		S10	S2	S9	T6	S1	S10
5		S6	T4	T4	C	S7	S7 |S9		S9	C	S10	C	S9	S2
6		T4	S7	S7	S6	T4	S9 |S7		S1	S9	S1	S10	S2	C
7		S9	S4	S9	S4	S6	S4 |S4		S2	S1	T5	S9	S10	S9
8		S7	S6	S6			T5 |T5			S10	S2	S1		S1
1	Draft 3	T2	C	S11	S10	S10	S5	Draft 4	T2	S3	S4	T1	T1	T7
2		S10	S10	C	T2	S8	T3		S4	S9	S8	S8	S4	S8
3		C	S8	S10	T3	C	T2		S3	S4	T1	S9	C	S4
4		S11	T3	T2	C	T2	S11		C	C	S3	S4	S8	T1
5		S8	S11	S8	S3	S11	C		T7	T1	C	T7	S9	T2
6		S3	S3	S3	S5	S3	S3		S9	S8	T7	S3	S3	S9
7		T3	S5	S5	S8	T3	S10		S8	T7	S9	C	T2	S3
8		S5			S11	S5	S8				T2	T2	T7	C
1	Draft 5	C	C	S11	T3	C	T3	Draft 6	C	C	S7	C	T1	S8
2		S11	T3	S7	C	S11	S7		S1	T1	S5	T1	S8	S10
3		S2	S2	C	S2	S2	S2		S10	S10	S10	S5	C	S7
4		T4	S7	S2	S11	T3	T7		T4	S8	C	S10	S1	S5
5		S7	S11	T7	T7	T4	S11		S8	S1	T1	S1	S10	T1
6		T7	T4	T4	S7	S7	T4		S7	S7	T4	S8	T4	C
7		T3	T7			T7	C		S5	T4	S8	S7	S7	S1
8										S5	S1		S5	T4
1	Draft 7	T5	C	S11	S3	T2	T5	Draft 8	C	C	S5	T7	S8	
2		C	T5	C	T5	C	C		T6	T3	T6	T3	T6	
3		T2	S3	T5	T2	S11	S4		S2	S8	T7	S8	C	
4		S3	S11	T2	S4	S6	T2		S1	S2	S8	C	S6	
5		S6	S6	S3	S11	S3	S3		S6	T6	S6	T6	S1	
6		S4	S4	S6	S6	S4	S11		S8	S6	C	S1	T3	
7		S11		S4	C		S6		T7	T7	S1	S5	T7	
8									T3	S1	S2	S2	S2	
9									S5	S5		S6	S5	

When analyzing the distribution within the tables, several tendencies can be discovered. Some of them directly relate to the RQ, whereas others are only tangentially related, yet still hold implications. For once, it can be seen that the rankings were never consistent across the different assessors. While for some papers at least some agreement existed regarding the best reviews (e.g., Draft 2), for others there were instances of the same review being rated best and worst by different people (e.g., Draft 3). In the case of Draft 3, this phenomenon is especially prevalent as shown by both, S5 and S11, holding the first and last place in the ranking depending on the evaluator. Thus, even though the sample size is rather small, this indicates that perceived review quality is highly subjective. Nevertheless, overall, some reviewers were better rated than others. Hereby, to little surprise, experience seems to play a relevant role. Within the group of teachers, *T1*, *T2*, *T5*, and *T6* are the most active when it comes to publishing scientific work and their reviews were usually placing rather well. In turn, *T3* has by far the most teaching experience but is not actively publishing. While their reviews were overall seen less favorably than the ones from the previous group, the students perceived them very positively. This could be a sign that the extensive teaching experience allowed them to provide feedback that was well crafted for the target audience. Yet, once again, the low sample size has to be pointed out. The reviews of *T4* and *T7*, however, did not stand out in comparison to the ones written by the students. Both teachers are mostly involved in project work and have neither extensive publishing nor teaching experience. Even though there was also some variation of the rankings assigned by the same assessor to the reviews of the same reviewer across the papers, this was overall rather stable, even more so for the rankings of the more experienced researchers.

Especially interesting with regard to the RQ is, however, the placement of ChatGPT within the rankings. Again, there was no unanimous sentiment. For the students and the teachers (also for both sub-groups) it got placements that ranged from first to last. Nevertheless, overall, it was seen favorably. For some papers, it was even pretty consistently ranked highly. This can be seen as another testament for ChatGPT’s capability to produce well-formulated and cohesive texts. It also supports the finding of [16] that ChatGPT is able to outperform humans when it comes to providing feedback to written texts. To our surprise, the students received the LLM-generated reviews worse than the teachers. Two of the students placed them last with one of them explicitly mentioning their suspicion that the review was generated by ChatGPT. Similar suspicions were also voiced by one of the teachers for another paper. Here, they mentioned two reviews as sounding ChatGPT-like. Yet, these were both provided by students and not the one that was injected for the experiment. Thus, the expectation is that ChatGPT was not involved in their creation. However, definitely determining if the students used ChatGPT is, of course, not possible.

As mentioned previously, in the course of the experiment, the participants were not only asked to provide their rankings but also to fill out a survey that covered several aspects. Similar to the reviews, the opinions were rather divided. Due to their different perspectives and roles within the course, the views of students and teachers will be discussed separately.

While two students did not see any positives, others lauded the structured feedback and the specific recommendations for improvement. Nevertheless, it was also highlighted that ChatGPT gives rather generic and “technical” feedback that does not fully grasp the intellectual essence of the papers. Further, it was criticized that some remarks were incorrect. When asked how big the influence of the ChatGPT reviews on the revision for the final paper was, the answers, again, varied. Some students completely ignored them or only incorporated them in a very limited fashion, while others valued them highly. In fact, two students described the impact of the ChatGPT-reviews on the final paper as higher than the ones by humans. However, one ranked them equal and four attributed (considerable) more importance to the human written reviews. Regarding the likelihood of using ChatGPT to generate reviews for their own future theses or seminar works as a foundation for improvements, all but the two students that were very critical of ChatGPT stated that they are (somewhat) inclined to use it.

Similar to the students, the teachers generally highlighted the clearness of the ChatGPT reviews as well as their strength in pointing out obvious flaws in the papers’ structure while criticizing their generic nature, the lack of depth with regards to the content and the susceptibility to incorrect statements. Naturally, the teachers had only very limited insight into the impact the ChatGPT reviews had on the final papers, especially in comparison to the reviews that were written by humans. Consequently, corresponding assessments could not be given with sufficient certainty. Further, while the students were asked how likely they are to use ChatGPT to provide them with feedback for future assignments, the question was adapted for the teachers. Instead, the teachers' stance on using ChatGPT reviews for future iterations of the just finished course and other courses was inquired, as well as their likelihood to recommend students to use ChatGPT to create reviews for theses and seminar works to get feedback to improve their writing. Within the scope of the course, all of them were positive towards incorporating LLM reviews in some capacity. It was especially pointed out how this could help the students with the structure and many rather general flaws of their texts, which, in turn, leads to better drafts and reduces the workload of the teachers for pointing these issues out, freeing them up to focus more on the intellectual aspects of the student papers. However, it was also noted that the students should be made aware when LLMs are used. Further, the importance of human feedback was emphasized several times. Thus, the sentiment was that using ChatGPT-reviews as an additional component was welcomed but these should not be used to completely replace human reviewers and the final say should always lie with the instructors, since these are considerably better at providing nuanced and detailed feedback. Moreover, the topics were provided by the teachers so they can give specific feedback regarding the expected scope, which is especially valuable during the initial phases of the writing process, where the foundations of the paper are laid.

From a general pedagogical perspective, it was pointed out that this could be a valuable chance to teach the students about opportunities and challenges of using LLMs and to very visibly illustrate their weaknesses when it comes to trustworthiness and their ability to go into detail about content, its meaning and implications, as well as to consider contexts. Besides supporting this specific course and maybe also dissuading some students from overreliance on ChatGPT for other courses within their studies, this would also prepare them for the time after they obtained their degree. After all, it is likely that many of them will be confronted with LLMs and potential application possibilities in their future jobs.

Additionally, several teachers pointed out that they consider further research in the field of incorporating LLM into teaching activities interesting and highly relevant. With a specific focus on the future use of ChatGPT in the conducted course, the exploration of different prompting approaches to improve the obtained reviews was suggested.

4. Discussion

As the experiment showed, LLMs can bring some value when used as reviewers to provide feedback for improvement. However, this mostly applies to rather general aspects such as structure or language, whereas specific and in-depth commentary on the content and methodology was lacking. Hereby, this impression was shared by both groups of participants, students and teachers. These findings are in line with current research that also points out that LLMs' current ability to provide meaningful critiques of research is rather limited [23]. Nevertheless, the involved teachers still overwhelmingly welcomed the idea of incorporating ChatGPT into future writing-assignments. Since the fundamental aspects of writing (scientific papers) are often lacking for the majority of IT students, addressing these takes up a significant portion of the teachers‘ time, which, in turn, is then missing for focusing on the actual content and methodology. Freeing these resources up could be a valuable relief, especially for institutions with rather limited means. While the teacher-to-student-ratio in the discussed course was very good, allowing for very intensive supervision, this is, by far, not the norm. Further, this opportunity was still not extensively used by most students, which might be due to aspects such as the required effort for attending consultations, scheduling conflicts with the students' other obligations (such as work), or the (misplaced) sorrow to be seen as bothering. An LLM solution, on the other hand, could be consulted independently, easily and flexibly in terms of time, reducing the potential obstacles mentioned above. Further, even though the low number of participants is a drawback of the experiment, the diverse composition of the participants adds significance to the findings. While the students were all master's students in the field of IT, withing this domain, they came from four different degree programs. Moreover, there were male and female participants, and the students came from seven different countries and varying education systems, decreasing the likelihood of corresponding biases in the overall results.

One thing that stood out in the results was how disparate the perception of the reviews was. While in some cases, certain tendencies emerged, in others, there was an extremely high variance. This applied not only to the reviews in general but also specifically to the assessment of the ChatGPT-reviews. This, in turn, also makes it challenging to create reviews that are positively received by all or at least most students and, at the same time, also seen as actually helpful from the teachers' perspective. Yet, to successfully establish LLM-support to students in courses like the discussed one, this should at least be aspired.

Theoretically, it would have been highly interesting to analyze if factors such as gender, origin, course of study, work experience, and especially the proficiency in the skills needed for the course (for instance determined by the grade received in the course as a proxy) influence the perception of the ChatGPT-reviews. However, for this a significantly higher number of participants would have been required, which would, besides all other challenges, not align well with the concept of the course. Potentially, conducting and aggregating numerous similar studies across many different institutions and countries might be an option, yet, assuring the comparability of the results might be challenging.

Apart from the low sample size, some other limitations also have to be considered when reflecting the results. For once, most of the students' papers were structured literature reviews, which plays into the identified strengths of ChatGPT as a reviewer. For works that, for instance, focus on theory-building, its weaknesses would most likely become even more apparent. Further, the focus was on IT students, who might be less experienced in writing than students from other faculties, making the basic feedback more valuable for them, which reflects in the perception. Presumably, the short time for incorporating the reviews into the draft (one week) also played a role, since extensive changes or adjustments to the general approach would not have been feasible anyways. Another limitation is that, theoretically, it could have been possible that students or teachers secretly used ChatGPT for writing their reviews, distorting the results. Finally, the task for the reviews was to provide feedback for improvement and not to provide an assessment of the quality, making the strong focus on aspects such as the structure more relevant.

Besides the actual experiment, out of curiosity, a very limited probe into the use of ChatGPT for the actual grading of the final submissions was also done but the results were very far from the grades that were determined by the instructors. However, this should not be overvalued, since this was explicitly not in the focus of this study and, thus, only a very basic prompt was used. It can be assumed that more sophisticated attempts could improve the performance considerably. However, even if this was the case, from an ethical and pedagogical perspective, using LLMs for actually grading such subjective tasks appears questionable. Yet, for other types of tasks, with less subjectivity involved (e.g., grading programming tasks or mathematical calculations), or as an uncomplicated and accessible self-check for students, these approaches could be very interesting.

5. Conclusion

To provide students with intensive support despite limited personal ressources, many higher education institutions are looking into harnessing LLMs as virtual tutors. In theory, this way, the teachers could focus on the more challenging problems and those that require human intervention, whereas others could be handled without needing their attention. To explore how such an approach for providing reviews for students‘ scientific writing assignments is perceived, an experiment was conducted. For the purpose of supporting students, the ChatGPT-reviews appeared to be a useful resource but the value in creating actual reviews for students or even real scientific conferences appears currently limited, due to the superficial nature of the provided analysis. For the future, testing other LLMs and varying prompting strategies (e.g., few-shot prompting) could be worthwhile approaches to improve the quality of the generated reviews. Further, an experiment where the students are not only given a LLM-review once, as was done here, but are actively and repeatedly incorporating such feedback throughout the whole writing-process could provide additional insights into the utility of such tools in the given context. Ultimately, if satisfying results regarding the review-quality can be achieved, a dedicated LLM-based review tool could be created and deployed to support students in independently refining their abilities in scientific writing.

References

1. Alipour, H., Pendar, N. and Roy, K.: ChatGPT Alternative Solutions: Large Language Models Survey, https://arxiv.org/pdf/2403.14469 [22.02.2025] (2024)
2. American Psychological Association: Ethical Principles of Psychologists and Code of Conduct: Including 2010 and 2016 Amendments, https://www.apa.org/ethics/code [01.05.2025] (2017)
3. Bloom, B.S.: The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educational Researcher 13, 4–16 (1984)
4. Brenner, A.M., Beresin, E.V., Coverdale, J.H., Louie, A.K., Balon, R., Guerrero, A.P.S., Roberts, L.W.: Time to Teach: Addressing the Pressure on Faculty Time for Education. Academic psychiatry : the journal of the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry 42, 5–10 (2018)
5. Brynjolfsson, E., Li, D., Raymond, L.: Generative AI at Work. National Bureau of Economic Research, Cambridge, MA (2023)
6. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al.: A Survey on Evaluation of Large Language Models. ACM Trans. Intell. Syst. Technol. 15, 1–45 (2024)
7. Du, H., Jia, Q., Gehringer, E., Wang, X.: Harnessing large language models to auto-evaluate the student project reports. Computers and Education: Artificial Intelligence 7 (2024)
8. Filippucci, F., Gal, P., Jona-Lasinio, C., Leandro, A., Nicoletti, G.: The impact of Artificial Intelligence on productivity, distribution and growth: Key mechanisms, initial evidence and policy challenges (2024)
9. Gallagher, H.L., Jack, A.I., Roepstorff, A., Frith, C.D.: Imaging the intentional stance in a competitive game. NeuroImage 16, 814–821 (2002)
10. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS quarterly, 75–105 (2004)
11. INSTICC: Templates, https://iceis.scitevents.org/Templates.aspx [19.02.2025] (2025)
12. Jääskä, E., Aaltonen, K.: Teachers' experiences of using game-based learning methods in project management higher education. Project Leadership and Society 3, 100041 (2022)
13. Kircher, T., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., van Os, J., Krach, S.: Online mentalising investigated with functional MRI. Neuroscience letters 454, 176–181 (2009)
14. Krecar, I.M., Kolega, M., Jurcec, L.: Perception of ChatGPT Usage for Homework Assignments: Students' and Professors' Perspectives. IAFOR J. Educ. 12, 33–60 (2024)
15. Lertzman, K.: Notes on Writing Papers and Theses. Bulletin of the Ecological Society of America 76, 86–90 (1995)
16. Li, J., Huang, J., Wu, W., Whipple, P.B.: Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. Humanit Soc Sci Commun 11 (2024)
17. Matthews, J., Volpe, C.R.: Academics' perceptions of ChatGPT-generated written outputs: A practical application of Turing's Imitation Game. AJET 39, 82–100 (2023)
18. Melo, C. de, Marsella, S., Gratch, J.: People Do Not Feel Guilty About Exploiting Machines. ACM Trans. Comput.-Hum. Interact. 23, 1–17 (2016)
19. OpenAI: Introducing ChatGPT, https://openai.com/index/chatgpt/ [16.02.2025] (2022)
20. Raiaan, M.A.K., Mukta, M.S.H., Fatema, K., Fahad, N.M., Sakib, S., Mim, M.M.J., Ahmad, J., Ali, M.E., Azam, S.: A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. IEEE Access 12, 26839–26874 (2024)
21. Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., Kilts, C.: A neural basis for social cooperation. Neuron 35, 395–405 (2002)
22. Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D.: The neural basis of economic decision-making in the Ultimatum Game. Science (New York, N.Y.) 300, 1755–1758 (2003)
23. Schryen, G., Marrone, M., Yang, J.: Exploring the scope of generative AI in literature review development. Electron Markets 35 (2025)
24. Simons, W., Turrini, A., Vivian, L.: Artificial Intelligence: Economic Impact, Opportunities, Challenges, Implications for Policy (2024)

25. Tafesse, W., Wood, B.: Hey ChatGPT: an examination of ChatGPT prompts in marketing. J Market Anal 12, 790–805 (2024)
26. Thelwall, M.: Can ChatGPT evaluate research quality? Journal of Data and Information Science 9, 1–21 (2024)
27. Visual Capitalist: Total market share of artificial intelligence (AI) tools worldwide in 2023, https://www.statista.com/statistics/1458132/ai-tool-market-share/ [01.05.2025] (2024)
28. Wang, Q., Gayed, J.M.: Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting. Computer Assisted Language Learning, 1–29 (2024)
29. Zhou, R., Chen, L., Yu, K.: Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In: International Conference on Language Resources and Evaluation, pp. 9340–9351 (2024)
30. Zoeten, M.C. de, Ernst, C.-P.H., Staegemann, D.: Perceptions of Annotated Explanations in Explainable AI: Humans versus ChatGPT. In: AMCIS 2025 Proceedings (2025)