Is Feedback In The Eye Of The Beholder? Chat Gpt Vs. Google Gemini In Giving Feedback On Students' Oral Presentation Scripts

Ksenija Benčina
Faculty of Humanities and Social Sciences,
J. J. Strossmayer University of Osijek, Croatia
E-mail address: kbencina1@gmail.com
ORCID: https://orcid.org/0000-0003-2705-0377

Lorna Dubac Nemet
Faculty of Medicine/Faculty of Dental Medicine and Health,
J. J. Strossmayer University of Osijek, Croatia
E-mail address: ldnemet@fdmz.hr/ldnemet@mefos.hr
ORCID: https://orcid.org/0009-0001-9606-1224

Abstract

Aim. The paper aims to demonstrate how integrating innovative technologies like Chat GPT and Google Gemini can offer a transformative approach to giving feedback to students.

Methods. This qualitative research explores the effectiveness of Chat GPT and Google Gemini in assessing students' oral presentation scripts. The sample consists of 12 oral presentation scripts written by medical students in years 5 and 6 for their Medical English class. The scripts are ordinarily evaluated based on the rubrics including clarity, cohesion, medical terminology, grammar, and syntax. Also, the Medical case report benchmarks are considered in the evaluation. The identical assessment criteria were used as prompts for both Chat GPT and Google Gemini and the feedback produced by each tool was then compared manually. The feedbacks were analysed for comprehensiveness, suggestions, problems and tone to determine which tool was more suitable for this task.

Results. The results indicate differences in the comprehensiveness and the focus of the assessment. Whereas Chat GPT relies for the most part on medical case report benchmarks and how they were/were not met by the students, Google Gemini analyses strengths, areas for improvement (clarity, concision, flow and transitions) and provides specific feedback by slide.

Conclusions. The results can contribute to a better understanding of the capabilities and limitations of Chat GPT and Google Gemini in providing feedback on students' oral presentation scripts. Furthermore, the findings can help educators select the most beneficial tool, with which they can provide personalised and immediate insights and thus support student learning and growth in medical English. Nevertheless, what is necessary for that potential to be adequately harvested with the lowest probability of error, bias and hallucination is quality-based prompt engineering and coaching on the part of the user.

Keywords: feedback, oral presentation scripts, Chat GPT, Google Gemini, medical case report

Introduction

Feedback is commonly defined as "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (Hattie & Timperley, 2007). Giving such information to students can be challenging as it is often time-consuming and requires a high level of expertise and individualised attention. Nevertheless, feedback presents a crucial part of the education process since it helps students recognise their strengths and weaknesses and identify areas of improvement. Figuratively speaking, feedback can be described as "the heart of learning" (Dubac Nemet, 2021), a powerful pump that sustains the entire teaching and learning process by functioning effectively. Despite its importance, research (Boud & Molloy, 2013) shows that feedback in higher education is still a contentious issue since many students express dissatisfaction with the type of feedback they receive and higher education institutions face more criticism for inadequacies in feedback than for almost any other aspect of courses. They also concluded that feedback in higher education is still very traditional and uninfluenced by practices and research on feedback from outside the education sector. Therefore, new approaches to feedback are necessary to effectively address these issues and ensure that it has a positive and lasting impact on learning.

Giving feedback to university students poses numerous challenges and addressing them requires a thoughtful approach (Haughney, Wakeman, & Hart, 2020). Their systematic literature review revealed that common challenges include subjectivity, the volume of feedback, clarity, timeliness, balancing positive and constructive feedback and student expectations. Also, based on our extensive experience as English language instructors, we have found that providing high-quality feedback on students' oral presentation scripts is a valuable, but a time-consuming task. Additionally, the unfavourable student-teacher ratio can make the feedback process extremely demanding. As a result, there is a growing interest in exploring technological solutions that can assist in providing feedback efficiently.

Artificial intelligence represents a technological breakthrough that has revolutionised almost all aspects of our lives. Large language models (LLMs), such as Chat GPT and Google Gemini, have emerged as powerful tools in the field of education. Chat GPT is described as a model that is trained to follow instructions in a prompt and provide a detailed response (openai.com/blog/chatgpt, 2022), while Google Gemini is trained to recognise and understand text, images, audio and more at the same time (Pichai & Hassabis, 2023). Generally speaking, the main difference is that at the moment Chat GPT is still primarily focused on text generation, whereas Google Gemini is a multimodal model that can process text, images and audio. It is important to note that both models use training data

up to the year 2022, which is especially relevant for the field of medicine where discoveries and new developments are made on a regular basis. The training data cutoff (2022) can potentially limit the feedback on medical topics that might have recent advancements.

The next sections present the research rationale, followed by methodology, results and discussion and conclusion.

Research

The aim of the research was to compare the effectiveness of free versions of Chat GPT -3.5 and Google Gemini 1.0 in providing feedback on students' oral presentation scripts. Determining the similarities and differences of both models' outputs can help teachers and students choose a more suitable model depending on their needs and make informed decisions to enhance the quality of teaching and learning.

This qualitative case study was carried out on a sample of oral presentation scripts provided by medical students in years 5 and 6 at The Faculty of Medicine at the J. J. Strossmayer University of Osijek. The scripts were written for their Medical English course in the academic year 2022/2023. For the most part the scripts were based on case reports from the Clinical Case Reports Journal published by Wiley Online Library and Oxford Medical Case Reports (OMCR) journal published by Oxford University Press. Both of the suggested sources are online, open access, peer-reviewed journals publishing original, educationally valuable medical case reports with the purpose of improving global health outcomes, therefore providing opportunities to cost-effectively augment medical education by harnessing the potential for self-directed learning, as well as scaffolded and training-based dealing with clinical practice cases. The students were asked to choose a case report of their own preference and to prepare an oral presentation following specific guidelines. The guidelines included ten Medical Case Report (MCR) benchmarks which served as a blueprint for oral presentations. The MCR benchmarks are a modified blend compiled from oral presentation instructions published by the University of Washington and Yale School of Medicine (McGee, n.d; Siegel, 2018). The oral presentation should start with the *Opener* which introduces the basic patient identifying information and chief concerns. The next benchmarks are related to the *History of the present illness*, *Review of systems*, *Past medical history*, *Family history* and *Social history* followed by a *Physical exam*, and *Results of laboratory testing, imaging and other* diagnostics. The oral presentation should end with a *Summary statement*, synthesising the important history and exam findings, framing the clinical problem and finally, leading the audience/listener into the *Assessment and plan.* In this way, a presentation script serves as a tool that helps students to organise the content in a meaningful way. It also ensures a structured presentation maintaining cohesion both within individual slides and across the entire presentation.

In order to address the specific criteria for the performance evaluation, the students were also introduced in advance to the rubrics including clarity, cohesion, the use of medical terminology, grammar and syntax. In that way, the feedback they receive functions as a feed-forward: a corrective instruction whose purpose is to provide the foundation for achieving a higher level of accomplishment, simultaneously triggering their evolution in the process of oral case presentation skills development. The presentation scripts, the starting point for preparing a PowerPoint presentation, help them to effectively deliver medical case reports in front of the audience, at the same time allowing the teacher to get insight into their presentation's blueprint during the preparation phase, making it possible to offer an intervention. It is important to note that this paper focuses solely on the initial preparation step, specifically on presentation scripts that should establish a strong foundation for actual oral presentations. In this process-driven activity, feedback is beyond doubt of utmost importance for the student, therefore it is crucial that they receive it as soon as possible during the pre-presentation phase. For this reason, we aimed to explore how integrating new technology can assist teachers in providing students with timely and constructive feedback during the initial phase of their presentation preparation.

Methodology

This case study was carried out on 12 randomly selected oral presentation scripts written by medical students. The selected scripts were written at the start of the 2022/2023 academic year, before Chat GPT and Google Gemini were widely available. The scripts were first anonymised and uploaded to both Chat GPT and Google Gemini. Both models received the same prompt for feedback: "You are a university medical English teacher. Provide personalised feedback on the student-written presentation script. Examine for clarity, cohesion, medical terminology, grammar, and syntax. Link to MCR benchmarks (Opener, History of the present illness, Review of systems, Past medical history, Family history, Social history, Physical exam, Results of laboratory testing, imaging and other diagnostics, Summary statement and Assessment and plan)".

The outputs of both models for each oral presentation script were downloaded and manually examined and analysed by a university English language instructor. The first step in the analysis was to determine the feedback structure provided by each model. A coding scheme was designed to categorise feedback outputs based on their general structure. This coding scheme employed a thematic approach, using three main codes: content, clarity, and engagement. The content code took into account the comments on the comprehensiveness of medical information and case presentation structure. The clarity code focused on the references to flow and transitions. Finally, the engagement code referred to suggestions for visuals, emphasis, delivery, etc. This thematic coding scheme enabled a systematic evaluation of the feedback provided by Chat GPT and Google Gemini resulting in a structured comparison of the two models' outputs.

Results and discussion

Tthe analysis of the feedback structure indicates that the two models had distinct approaches and emphasized different areas. Chat GPT focused on the ten MCR benchmarks and provided separate comments for each. A general comment on the oral presentation script was given at the end of the output, summarising the key points of the detailed comments related to MCR benchmarks. On the other hand, the outputs by Google Gemini have a more elaborate structure starting with strengths and areas for improvement organised in bullet points followed by specific feedback for each slide. At the end of the output, Google Gemini also listed a few additional tips and summarised the comments.

When it comes to detailed feedback for each MCR benchmark or slide, it can be observed that Chat GPT had a more thorough and descriptive approach, whereas Google Gemini suggested which information should be emphasised or explained. The following example illustrates these two different approaches:

"Physical Exam: The physical examination findings are well-described and highlight key clinical indicators of a potential ectopic pregnancy, such as hypotension, tachycardia, abdominal distension, and cervical motion tenderness. The presentation effectively conveys the urgency and severity of the patient's condition". (Chat GPT, Oral presentation script 4, 2024)

"Physical Exam: Highlight the key findings: hypotension, abdominal distension and tenderness, positive cervical motion tenderness". (Google Gemini, Oral presentation script 4, 2024)

Besides this, differences were also observed in the comments related to the content presented on each slide. Chat GPT's approach is more analytical as it comments on what is presented well, but also in some cases suggests what should be added to provide more context.

"The physical examination findings are clearly presented and provide important information about the patient's vital signs and abdominal examination. However, additional details about other relevant physical examination findings could enhance the completeness of the presentation". (Chat GPT, Oral presentation script 1, 2024)

On the other hand, Google Gemini's comments did not include any information about the lacking content, but they simply reiterated the information for each slide mentioning what should be focused on, discussed, etc.

"Briefly discuss differential diagnoses and how they were ruled out. Emphasise the severity of the lactic acidosis despite normal vital signs". (Google Gemini, Oral presentation script 1, 2024)

Furthermore, it was also noted that there were differences in the number of comments given regarding clarity. Whereas Chat GPT suggested only in two feedback outputs that the content *could be structured more cohesively* and *organised more systematically*, Google Gemini provided more detailed comments on clarity in all feedback outputs (e.g. *combine information, consider using a bulleted list, some sections can be more concise, improve transitions between sections by summarising key points at the end of each section*).

Google Gemini also focused more on the engagement of the audience as it suggested in all feedback outputs which images should be incorporated (*consider incorporating an image of a fallopian tube or an ultrasound image)* as well as that the presenter should consider the audience's medical background and adjust the level of medical terminology accordingly. Chat GPT suggested including visuals only in three feedback outputs (*including images from the ultrasound and CT scan could enhance visualisation and understanding of the diagnostic findings*).

What we found particularly worth further research and reflection was the fact that even though Google Gemini listed grammar and syntax in its feedback elements, in the majority of the outputs it either did not provide accurate feedback or in some cases failed to comment on obvious issues with spelling, sentence structure and connections, tenses, etc. Google Gemini remained silent for the following student inputs, with obvious mistakes:

"Sister is diaagnosed with hypertension and had glaucoma for which she has routinely check-ups" (Oral presentation script 3). This sentence has numerous mistakes such as a missing possessive pronoun, misspelt words, the use of inconsistent verb tenses and mixing up adverbs and adjectives.

The next example shows that some sentences written by students had a faulty sentence structure: "The patient was ruled out diagnoses such as /*The patient's rule-out diagnoses were*/ myocardial infarction (because of normal troponin), acute pyelonephritis (because of normal creatinine and urea) and acute appendicitis (due to the presence of epigastric pain)". (Oral presentation script 12)

These possible omissions indicate that we still cannot expect LLMs to be completely accurate and reliable and teachers should carefully validate such outputs.

Limitations

The growing influence of Artificial Intelligence in various fields inspired us to explore how LLMs can assist teachers in providing feedback. However, using LLMs in education remains a relatively new and underexplored area, requiring careful consideration of potential limitations. This study was carried out on a relatively small sample of students' oral presentation scripts, which may limit the generalisability of the findings to the broader population of medical students. Additionally, the increasing availability of specialised LLMs and AI tools has made it challenging to determine whether students utilised them while preparing their presentations. Thus, the sample for this research included scripts written at the beginning of the 2022/2023 academic year, when LLMs were not yet widely accessible. Future studies should take into account students' growing reliance on digital tools.

Another limitation is that this study focused only on one aspect of the presentation, whereas other elements included in the overall assessment process – such as STRUCTURE AND ORGANISATION, CONTENT and DELIVERY – were not examined in detail. According to Dubac Nemet (2021, p. 68), these elements are essential for a comprehensive evaluation of oral presentations. Within the

scope of STRUCTURE and ORGANISATION, students' presentations and students as presenters were evaluated based on their implementation of the structural framework, slide layout, and bullet point design. Furthermore, their performance was scrutinised for clarity, coherence, and proper use of signposting. In the CONTENT category, the key elements included the use of professional terminology, successfulness in addressing the target audience (professional level) and including adequate use of references. Finally, within the DELIVERY category, the presenters were assessed based on their fluency, pronunciation of general English and medical English terms as well as the vocal aspects of delivery, focusing on their command of rhythm and stress (pacing, articulation, pitch, speed, volume).

While this study provides insights into the potential of LLMS for providing feedback on oral presentation scripts, it is important to consider the broader contextual challenges in medical education. The increasing number of students enrolled in medical programmes has led to alarmingly unbalanced student-teacher ratios. For instance, at the Osijek Faculty of Medicine, student enrolment (per academic year) has nearly doubled: from 40 in the year 2000 to around 75 in recent years. This has significantly reduced the amount of time teachers can dedicate to each student, raising the questions about traditional feedback methods. Although LLMs and AI tools could serve as potentially cost-effective solutions to this problem, further research is needed to evaluate their effectiveness in addressing these challenges.

Conclusion

Nowadays incorporating large language models such as Chat GPT and Google Gemini into the education process seems inevitable. Such models can offer innovative approaches to teaching and learning especially when it comes to providing personalised and timely feedback to students. Additionally, these models have the potential to support problem-solving, critical thinking, and knowledge acquisition in ways that traditional methods may not be able to achieve, at least not in the time frame of LLMs' reaction, as well as its outreach and flexibility. The ability of Chat GPT to perform at a level equivalent to a passing score for a third-year medical student on the USMLE exams (Gilson et al., 2023) mirrors the potential for providing immediate feedback and support in evaluating students. Nevertheless, what is necessary for that potential to be adequately harvested with the lowest probability of error, bias and hallucination is quality-based prompt engineering and coaching on the part of the user. Furthermore, updated versions and models designed specifically for medical applications (trained on medical data) will also significantly contribute to trustworthiness and accountability.

With the limitation of it being a pilot study, with a sample size of 12 presentation script feedbacks received from Chat GPT and Google Gemini, our results may, to a certain extent, contribute to a better understanding of the capabilities

and limitations of Chat GPT and Google Gemini in providing feedback on students' oral presentation scripts. Consequently, these findings can help educators select the most beneficial tool to improve feedback for oral presentation scripts and assist in advancing students' self-directed learning. By using the analytical capabilities of Chat GPT and Google Gemini together with educators' fine-tuning (RLHF-reinforced learning through human feedback)(Stiennon, 2020), personalised and immediate insights can be provided to students and in that way support their learning and growth in medical English.

Training students on appropriate prompt engineering as well as teaching them about the importance of validating LLMs' outputs with reputable resources and raising awareness of the importance of data privacy in both digital and non-digital settings should by all means become a part of the medical curriculum for the next generation of healthcare professionals to be prepared for responsible and efficient patient care.

References

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment and Evaluation in Higher Education, 38*(6). 698-712.

Dubac Nemet, L. (2021). The importance of assessment in task-based Medical English courses for the evolution of students' presentation skills. *5. međunarodna konferencija "Suvremeni izazovi u poučavanju jezika struke"*, 65-77.

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A. & Chartash, D. (2023). How does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education, 9,* e45312. doi: 10.2196/45312.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1). 81-112.

Haughney, K., Wakeman, S., & Hart, L. (2020). Quality of Feedback in Higher Education: A Review of Literature. *Educational Sciences*, *10*(3). 60.

McGee, S. (n.d.) *Oral Case Presentation.* Retrieved from https://uw.pressbooks.pub/fcmtextbook/part/oral-case-presentation

OpenAI. (2022). *Introducing Chat GPT*. Retrieved from https://openai.com/index/chatgpt/.

Pichai, S., & Hassabis, D. (2023). *Introducing Gemini: our largest and most capable AI model*. Retrieved from https://blog.google/technology/ai/google- gemini-ai/#capabilities.

Siegel, M. D. (2018). *The Yale Way.* Retrieved from https://files- profile.medicine.yale.edu/documents/eccf7de4-4c16-4f56-9a91-59b8015e48da

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to summarize with human feedback. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems* (NIPS'20). Curran Associated Inc., Red Hook, NY, USA, Article 253: 3008-3021. Retrieved from: https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf

PROJECT AND PROFESSIONAL PRACTICE