Article types: original papers

Title: Effectiveness of a Large Language Model-Based Feedback System for Case Report Writing in Novice Rehabilitation Staff Education: A Mixed-Methods Study

Authors:

Yuuto Tonouchi(OT)[1]

Shunsuke Nakai(OT/MS)[1,2]

Kayo Nurakami(PT)[1]

Yuki Kataoka(MD/MPH/DrPH)[3,4,5,6,7]

Author Affiliations:

[1] Department of Rehabilitation, Kyoto Min-iren Asukai Hospital, Kyoto, Japan; [2] Osaka Metropolitan University Graduate School of Rehabilitation Science; [3] Department of Internal Medicine, Kyoto Min-iren Asukai Hospital, Kyoto, Japan; [4] Section of Clinical Epidemiology, Department of Community Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan; [5] Department of Healthcare Epidemiology, Graduate School of Medicine and Public Health, Kyoto University, Kyoto, Japan; [6] Department of International and Community Oral Health, Tohoku University Graduate School of Dentistry, 4-1, Seiryo-machi, Aoba-ku, Sendai, Miyagi, 980-8575, Japan; [7] Scientific Research Works Peer Support Group (SRWS-PSG), Osaka, Japan;

Corresponding Author:

Yuuto Tonouchi

Kyoto Min-iren Asukai Hospital

Department of Rehabilitation

89, Tanakaasukai, Sakyo-ku, Kyoto-city, Kyoto, 606-8226

Tel: +81-75-712-9091

Fax: +81-701-9183

Email: yuuto_12_ru_02@yahoo.co.jp

Abstract

Objectives: To develop a large language model (LLM) based feedback system to improve the efficiency of case report writing in novice rehabilitation staff education.

Design: A sequential mixed methods study.

Methods: We conducted a preliminary survey to identify burdensome feedback tasks and developed prompts using the Claude 3 Opus. We implemented the feedback system with Google Apps Script and Slack chatbots. Effectiveness and usability were evaluated through surveys. The study included five novice rehabilitation staff who joined our hospital in April 2024.

Results: All novice staff reported that the LLM feedback was equivalent to previous human feedback and helpful for their learning. The System Usability Scale (SUS) scores showed high usability (median: 90, range: 70-95). Three instructors (60%) agreed the system saved time and reduced guidance sessions, while four (80%) felt it would alleviate their future burden. However, opinions varied regarding the feedback content's suitability and its potential to enhance novice staff learning outcomes.

Conclusion: The LLM-based feedback system for case reports showed potential to reduce instructors' burden and provided an efficient learning environment for novice rehabilitation staff. Future research should focus on system revision and further evaluation.

This study was pre-registered in the UMIN Clinical Trials Registry (UMIN-CTR) (Trial ID: UMIN000053315). https://center6.umin.ac.jp/cgi-bin/icdr/ctr_reg_list.cgi

Key word:

large language model, education, case report, feed back

Introduction

Effective postgraduate education for novice staff in healthcare institutions is an important issue. During the orientation period, many novice staff experience positive emotions, but also feel burdened by anxiety about new clinical experiences, excessive responsibility, and physical labor, facing many stresses as care providers.[1,2] To address these issues, online learning has received attention as an alternative to traditional training methods,[3] which is suitable for new graduate healthcare professionals both at the orientation stage and as a follow-up learning method.[4]

In Japan, postgraduate education for healthcare professionals has been standardized. For nurses, the revised guidelines for new nurse training[5] and the learning support book[6] show educational goals and indicators to ensure the quality of nursing practice skills in all facilities. For physiotherapists, the first edition of the "New Physiotherapist Staff Training Guidelines" was issued on 1 November 2020, showing goals for attaining clinical practice skills and systematizing educational guidelines.[7] For occupational therapists, standardized educational guidelines have not been presented.[8] The educational program for novice rehabilitation staff is determined by each facility's practice. During the first year of postgraduate education, facilities often assign a case report creation as part of training, similar to undergraduate education.[9]

The feedback system for these reports is underdeveloped. A survey of 58 rehabilitation facilities in Tokyo revealed that among the 48 facilities providing novice rehabilitation staff education, only about 40% have well-developed systems, and many rely on instructors with less than 10 years of experience, who often guide new employees based solely on their own experience.[10] Additionally, 17 out of 58 facilities (35.4%) could not offer novice education during working hours, reflecting inadequate instructional time and increasing the burden on instructors.

Enhancing instructional efficiency can save time and resources and support effective work processes. The development of large language models (LLMs) has advanced natural language processing technology, offering new opportunities to improve education and work processes. Studies have highlighted the use of ChatGPT for feedback on student reports and peer review of research papers.[11,12] In Japan, products using LLMs have already been developed to pre-read contracts in the legal field.[13]

However, there are no reports on the use of LLMs for providing medical feedback on case reports prepared by novice rehabilitation staff. This study aims to develop a system that uses LLM-based inference[14] to improve feedback efficiency on case reports in the education of novice rehabilitation staff.

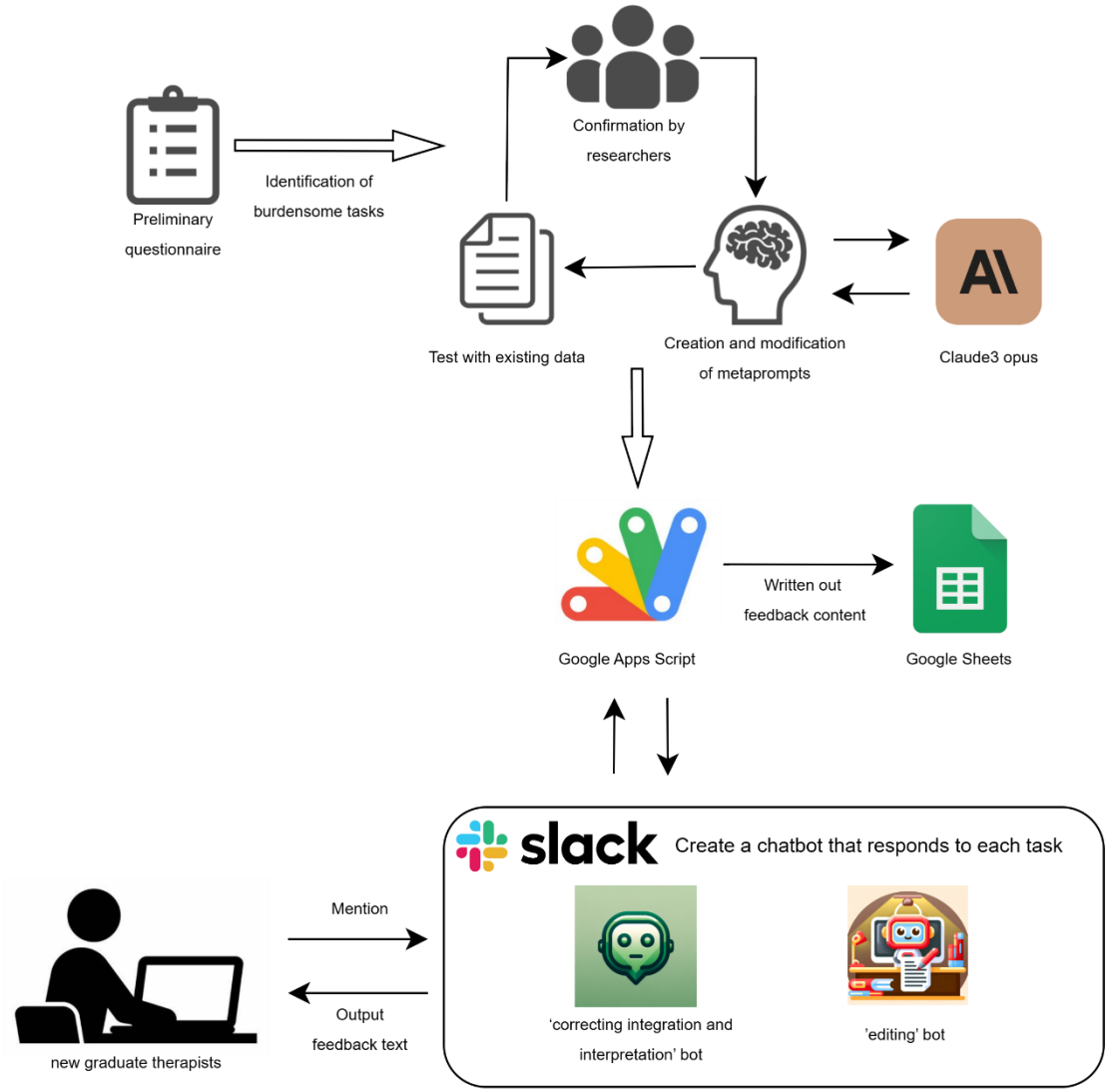

Figure 1: Overview

Methods

Study design and ethical considerations

This study is a sequential mixed methods study[15] targeting novice rehabilitation staff. Specifically, we combined both qualitative and quantitative methods and adopted a sequential process, using findings from each stage to inform the next, to develop a new education system tailored to the specific needs of novice rehabilitation staff education.

We conducted this sequential mixed methods study in five stages. In the first stage, we conducted a questionnaire survey among staff with experience in novice education at our hospital to identify the tasks that were burdensome when providing feedback on case reports prepared by novice rehabilitation staff. In the second stage, we created meta-prompts for each identified task using existing case reports. A meta-prompt is a text used to provide instructions to an LLM during a conversation with the LLM.[16] In the third stage, we developed a chatbot system that responds based on the completed meta-prompts. In the fourth stage, we introduced the LLM feedback system experimentally to novice rehabilitation staff. In the fifth stage, we conducted a questionnaire survey among novice rehabilitation staff and instructors to evaluate the effectiveness and convenience of this system. (Figure 1)

The study protocol was approved by the Ethics Committee of our hospital (ID: 2024-0502). We obtained individual informed consent from the participating staff members. This study was also pre-registered in the UMIN Clinical Trials Registry (UMIN-CTR) (Trial ID: UMIN000053315). To report this study, we followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines for cohort studies (supplemental etable 1).[17]

Overview

Stage 1) Survey of tasks that are burdensome for educational instructors

We conducted a questionnaire survey among 29 rehabilitation staff members who had experience in novice education within the past five years. Using Google Forms, we identified the average time required for providing feedback on case reports and the tasks that were burdensome when making corrections and providing guidance.

Stage 2) Create meta-prompts for each task

We created meta-prompts based on the content obtained from the preliminary questionnaire. We verified the appropriateness of the meta-prompts through visual confirmation by three of the authors: two occupational therapists and one physiotherapist (with 8-14 years of experience) who had experience in novice education. We repeated the process until we obtained appropriate meta-prompts. We used the "Claude 3 Opus" API from Anthropic.[18]

Stage 3) Implementation of LLM feedback system

We used Slack as the implementation environment. Slack is a cloud-based platform for business communication.[19] We created chatbots to respond to each task of "correcting integration and interpretation" and "editing" based on the meta-prompts developed in the second stage. We developed the backend of the chatbots using Google Apps Script (GAS). Subsequently, we set up dedicated channels for each novice rehabilitation staff member, enabling them to receive feedback by sending case reports as messages to the dedicated chatbots for each task within that channel (Figure 2). To follow the Act on the Protection of Personal Information,[20] we instructed novice rehabilitation staff on handling patient personal information as stipulated in case reports.

Stage 4) Introduction of LLM feedback system

The study included five novice rehabilitation staff (three physiotherapists and two occupational therapists) who joined the rehabilitation department of our hospital in April 2024. They used the newly developed 'LLM Feedback System' for their case reports, which they prepared during the first term (April to June, a three-month period) of our hospital's novice rehabilitation staff education curriculum (a seven-month program from April to October). The case reports summarized the evaluation and treatment planning for one patient each staff member was responsible for.

Stage 5) Evaluation

We conducted a questionnaire survey among novice staff and instructors. For instructors, we asked questions regarding the "comparison of instruction efficiency between the conventional method and using a large language model" and "whether the feedback content was appropriate". For novice rehabilitation staff, we asked questions regarding "whether the feedback content was appropriate" and the "System Usability Scale (SUS)". We summarized with descriptive statistics. In addition, we conducted content analysis of the post-questionnaire using an educator's lens.

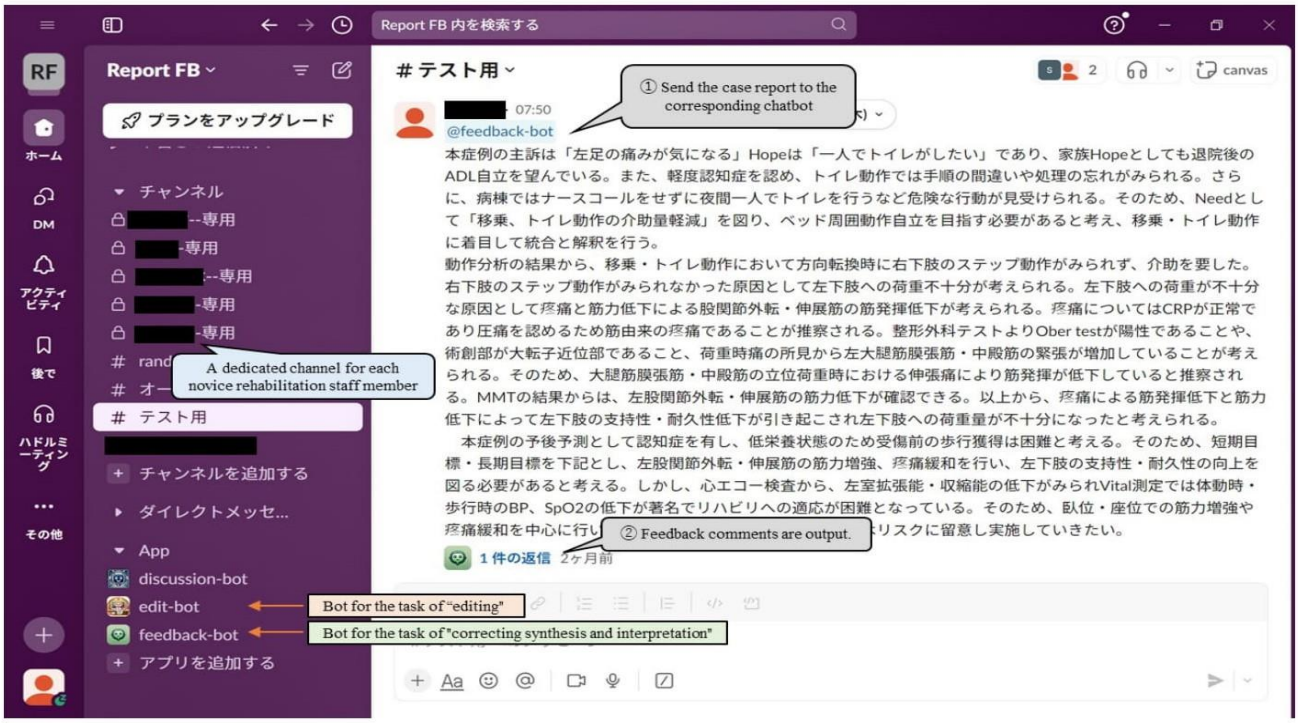

Figure 2: The actual operating screen

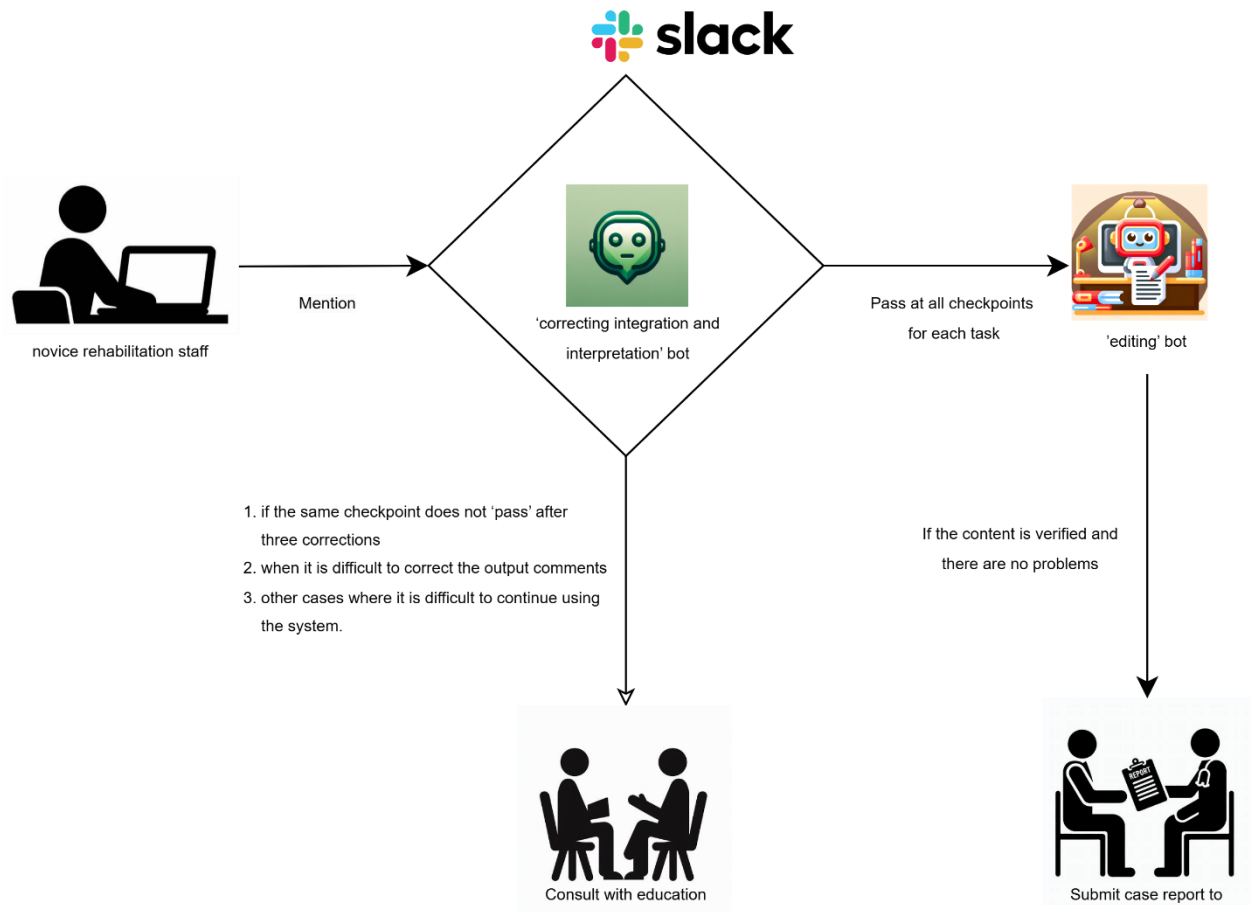

Figure 3: Flowchart for using the system

Results

1) Survey of tasks that are burdensome for educational instructors

The response proportion for the pre-survey was 20 out of 29 (69%). The years of experience of the novice education instructors were as follows: 4 (20%) with 2-5 years, 7 (35%) with 6-10 years, and 9 (45%) with 11 years or more. Over 80% of the staff responded that it takes 30 minutes or more for a single revision and feedback session. The tasks that require the most time for revision were: (1) the integration and interpretation process (90%); and (2) the discussion process (95%) (Table 1). "Integration and interpretation" is the process of considering the causal relationship between the subject's functional level and rehabilitation assessment results to formulate a hypothesis.

Table1. The pre-implementation survey results

	n	%
How many years of experience do you have?		
2-5 years	4	20
6-10 years	7	35
11 years or more	9	45
How long does it take for one correction and feedback session? (This includes all work such as reading, correcting, and instructing of reports.)		
15 minutes or less	1	5
15-30 minutes	3	15
30-45 minutes	7	35
45-60 minutes	6	30
60 minutes or more	3	15
What tasks are time-consuming in terms of corrections and feedback? (Multiple answers allowed)		
Introduction	3	15
Case presentation	1	5
Rehabilitation assessment	5	15
Integration and interpretation	18	90
Rehabilitation goal setting	10	50
Discussion	19	95
Literature search	4	20
Overall presentation and structure	9	45

2) Create meta-prompts for each task

We created meta-prompts corresponding to the tasks of "correcting integration and interpretation" and "editing." We designed these meta-prompts to output correction results regarding the "overall structure and writing style" of each task, based on existing case reports.

Please refer to the details of the latest version of the meta-prompts published on GitHub.
(https://github.com/youkiti/report-feedback)

In the meta-prompt for "correcting Integration and Interpretation," we set it to perform revisions in the order of the following checkpoints 1-5. (1. Explanation of the case (Is the overall picture of the case concisely explained at the beginning?), 2. Focus point (Describe the focus point in this report and the reason for it), 3. Problems with the focus point (Break down the focus point from step 2 and specifically describe which process is considered problematic), 4. Comparison with physical therapy or occupational therapy evaluation (For each of the issues mentioned in step 3, make a description that compares them with the actual physical therapy or occupational therapy evaluation results), 5. Key points necessary for improvement (Describe the "rehabilitation treatment program" and "prognosis prediction for the case")) In addition, each checkpoint is judged as either "passed" or "requires revision." Even if it is 'passed', comments for further improvement will be provided to enhance the quality of the report. We focused the criteria for pass/revision not on the quality of the specialized content, but on the goal of being able to write with the minimum necessary sentence structure as a preliminary stage before submitting to the instructors. The meta-prompt for "editing" was created by modifying some of the content of the "Prose polisher" in the Prompt Library published by Anthropic.[21]

3）Implementation of LLM feedback system

First, novice rehabilitation staff send their prepared case report (integration and interpretation text) to the chatbot for "correcting integration and interpretation" to receive feedback. If they pass all the checkpoints for each task, they then send the case report to the "editing" chatbot. Afterwards, the completed case report is submitted to each educational instructor (Figure 3). However, if any of the following apply while receiving feedback from the chatbot for "correcting integration and interpretation": 1) the same checkpoint is not "passed" even after three revisions, 2) it is difficult to

make corrections based on the chatbot's comments, or 3) it is difficult to continue using the system for any other reason, they consulted with the instructor even if not all checkpoints are passed.

4) Introduction of LLM feedback system

All five novice staff utilized the LLM feedback system. At the stage of submitting their first report, four out of the five passed all the checkpoints of the "correcting integration and interpretation" bot. As for the usage situation per one cycle, they submitted to the instructor after the "correcting integration and interpretation" bot provided feedback once or twice and then the "editing" bot provided feedback once. Three of them also used the LLM feedback system in the same manner even after the instructor provided feedback.

Including the usage after the instructor provided feedback, the number of times each chatbot was used for one case was as follows. The "correcting integration and interpretation" bot provided a median of 3 feedback sessions (range 1 to 4). The "editing" bot provided a median of 3 feedback sessions (range 1 to 3).

5) Evaluation

The response rate for the post-implementation survey was 5 out of 5 (100%) for both instructors and novice staff. The scoring for each question is explained as follows: 'Strongly Disagree' is 1 point, 'Disagree' is 2 points, 'Neutral' is 3 points, 'Agree' is 4 points, and 'Strongly Agree' is 5 points.

In the Survey on the Appropriateness of Feedback Content from LLM for instructors, for the question 'Comparable to past human feedback?', the median score was 3 points (range 2 to 4). For the question 'Helpful for novice staff learning and growth?', the median score was 3 points (range 1 to 4). In the Survey on the Efficiency of Instruction for instructors, for the question 'Did using this system save time and reduce the number of instruction sessions compared to traditional human-only feedback?', the median score was 4 points (range 1 to 5). For the question 'Do you think using this system will reduce the burden on instructors in the future?', the median score was 4 points (range 1 to 5). For the question 'Do you think using this system will improve the learning efficiency of novice staff in the future?', the median score was 3 points (range 1 to 4).

In the Survey on the Appropriateness of Feedback Content from LLM for Novice staff, for the question 'Comparable to past human feedback?', the median score was 4 points (range 4 to 5). For the question 'Helpful for novice staff learning and growth?', the median score was 5 points (range 4 to 5). In the Survey on the System Usability Scale (SUS) for Novice staff, the median SUS score was 90 points (range 70 to 95). (Table 2)

For the results of the free-text comments, please refer to Table 3 and Table 4.

Table2. The Post-implementation survey results

	Instructors	Novice staff
	Median (Range)	
1. Survey on the Appropriateness of Feedback Content from LLM		
1) Comparable to past human feedback?	3 (2 to 4)	4 (4 to 5)
2) Helpful for novice staff learning and growth?	3 (1 to 4)	5 (4 to 5)
2. Survey on the Efficiency of Instruction		
1) Did using this system save you time and reduce the number of instruction sessions compared to traditional human-only feedback?	4 (1 to 5)	
2) Do you think using this system will reduce the burden on instructors in the future?	4 (1 to 5)	
3) Do you think using this system will improve the learning efficiency of novice staff in the future?	3 (1 to 4)	
3. Survey on the System Usability Scale (SUS)		
1) I think that I would like to use this system frequently		4 (4 to 5)
2) I found the system unnecessarily complex		1 (1 to 2)
3) I thought the system was easy to use		4 (4 to 5)
4) I think that I would need technical support to use this system		3 (1 to 3)
5) I found the various functions in this system were well integrated		4 (4 to 5)
6) I thought there was too much inconsistency in this system		2 (1 to 2)
7) I would imagine most people would learn to use this system quickly		5 (4 to 5)
8) I found the system very cumbersome to use		1 (1 to 2)
9) I felt very confident using the system		4 (3 to 5)
10) I needed to learn a lot of things before I could get going with this system		2 (1 to 3)
The calculated SUS score		**90 (70 to 95)**

Explanation of question scores: 1: Strongly Disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly Agree
Abbreviations: LLM, Large-scale Language Models;

Table 3. The Post-implementation survey results from instructors (Free Text Comments)

	Positive comments	Negative comments
1. Survey on the Quality of Feedback Content from LLM	-If used effectively, there would likely be fewer omissions in each content. -Basic text is fine, but it seemed unable to express the flow of the story, especially the layered logical thinking. -There are methodologies for constructing explanatory text, but sensitivity to words and the skill in handling words are not singular; in that sense, there is no single correct answer. Both machines and humans have their strengths and weaknesses. -It seems that getting approval from AI is boosting confidence. I think this could potentially become a driving force for overall progress.	-I felt that providing feedback on narrative content is difficult for AI. -I think it was unable to fully interpret observed phenomena such as gait analysis. In that regard, it's necessary to delve deeper using the PDCA cycle based on hypotheses, so AI's knowledge and definitions alone are insufficient. -It seemed that it wasn't being utilized to its full potential. -While it's also influenced by the abilities and qualities of new staff, there was an overreliance on literature-based content, leading to a lack of focus on observing and thinking about actual patients. -While this concern may become unnecessary if it becomes the norm in the near future, I worried that it might hinder the development of skills in writing one's own text and in mentoring juniors and students.
2. Survey on the Efficiency of Instruction	-By the time the report was submitted, the basic structure and writing had already been somewhat revised, allowing us to focus on providing guidance on more advanced aspects. -I believe the efficiency of guidance improved because not only were the reports submitted with spelling errors, typos, and sentence structures already organized to some extent, but we were also able to review the Slack interactions before the final submission. -The need for guidance on basic Japanese syntax has definitely decreased. However, it's highly likely that the individual abilities of the novice staff also play a role in this improvement. -In recent years, whether due to changes in practicum guidelines or the impact of COVID-related restrictions on practical training, I feel that more novice staff members struggle with articulating their thoughts in writing compared to before. When submissions have insufficient grammar and structure, it's challenging to interpret them. With these points already addressed, I felt we could	-While there were no concerning issues with individual sentences, the overall impression was that of an immature writing style due to the lack of variety in conjunctions. It was necessary to suggest changing the conjunctions to better suit the context. -Since the content produced by novice staff was initially disorganized, we first had them submit reports in bullet-point format. As a result, the efficiency of guidance remained unchanged from before. -Setting aside the debate of whether it's good or bad, both instructors and novice staff felt that since AI judged it as passing, major revisions might not be necessary. This allowed them to complete the feedback within the designated time. -The comments from the AI often aimed to add more information to make the details clearer. However, since we specifically asked for focused and summarized reports, it seemed that the AI's comments were not utilized much.

	smoothly transition into providing feedback on clinical reasoning. -The fact that feedback can be received even while at home -The specific guidance on the text written by the novice staff seemed to help with their learning and presentation preparation.	-Ultimately, it is up to the user. Some people may improve their learning efficiency by using it, while others might mistakenly believe they are improving. -It was unclear whether the new staff understood how to use it effectively. -The basic thoughts and considerations in the submitted assignments were fine, but it seemed they were overly influenced by the feedback content.
3. Did the use of this system change the content of the feedback you provide?	-In terms of summarizing content, it was helpful in some aspects of how to explain things. -I was able to focus on providing specialized feedback. -It seems that efficiency improved and time was saved, allowing for more thorough consideration of the presentation content. -Basically, I was able to focus on providing specialized feedback.	-Compared to previous novice staff training, I don't think there's much difference at this point. -It may have been a hindrance in encouraging the thinking process of observing and considering the patient.
4. Are there any areas where the system needs improvement?	-I am humbled. Is it the richness and difficulty of the freedom in Japanese expressions? -More than the system itself, I felt that it is necessary to not only explain at the beginning but also to discuss and understand how to use it along the way. -Among the comments from the AI, there were some that praised the writing unnecessarily. I felt that if the AI approves but the instructor thinks it should be corrected, the beginner staff might get confused. - I think there could be more feedback that encourages thinking.	

Abbreviations: LLM, Large-scale Language Models;

Table 4. The Post-implementation survey results from novice staff (Free Text Comments)

	Positive comments	Negative comments
1. Survey on the Quality of Feedback Content from LLM	-The assistance with text structure was invaluable, especially for someone like me who struggles with writing. -It was helpful that you pointed out areas needing improvement or additions. -I received feedback on detailed aspects like interpretation and reasoning. -Clearly indicating strengths and areas for improvement was useful. - I found it helpful in summarizing reports in fields where I have limited knowledge, as it provides expert opinions based on specialized knowledge. -For novice staff who lack specialized knowledge, I found it particularly useful in helping them articulate their thoughts into written form, even when their ideas are clear in their minds but difficult to express in writing. -Responses come back quickly. -The feedback is documented, so it's not forgotten. -It reduces mental stress. -It's beneficial because detailed feedback is provided, even on minor points.	-I appreciated the additional advice. However, more specific expert opinions would have been beneficial.
2. Are there any areas where the system needs improvement?	-I thought that if the feedback could provide more insightful opinions, it would result in an even more robust report. -I felt it was a bit difficult to submit casually since everyone can see it. I think it would be more user-friendly if we could use it easily and immediately after organizing our thoughts. -I think it would be more readable if there were underlines or similar markings just for the points that need improvement.	

Abbreviations: LLM, Large-scale Language Models;

Discussion

In this study, we developed and evaluated a feedback system using LLM for case reports written by novice rehabilitation staff. All five novice staff used the LLM feedback system. They submitted their reports to the instructor after receiving feedback from the "correcting integration and interpretation" bot once or twice, followed by feedback from the "editing" bot once. Additionally, at the stage of submitting their first report, four out of the five novice staff passed all the checkpoints of the "correcting integration and interpretation" bot. Post-implementation surveys received a 100% response proportion from both novice staff and instructors. All five novice staff members reported that the LLM feedback was equivalent to previous human feedback and helpful for novice staff learning and growth. The System Usability Scale (SUS) scores showed high usability. Instructors had varied opinions on the appropriateness and efficiency of LLM feedback. Three instructors (60%) agreed that the system saved time and reduced the number of guidance sessions, whilst four (80%) felt the system would alleviate the burden on instructors in the future. However, divergent perspectives emerged regarding the suitability of the feedback content and its potential efficiency in enhancing the learning outcomes of novice staff.

The system developed in this research has the potential to reduce the burden on educational instructors and contribute to providing an efficient educational environment for new rehabilitation staff. Firstly, regarding the reduction of burden on educational instructors, previous studies on student education using LLMs have reported a reduction in teachers' workload and stress, as well as improved instructional efficiency.[22-24] In the survey results of educational instructors in this study, 80% of instructors responded that they "think it will contribute to reducing the burden on instructors in the future", and positive free comments included "reduced guidance on overall text composition" and "able to focus on specialized feedback". These results suggest that the promptness and convenience of feedback provided by this system may allow educational instructors to allocate more time to specialized instruction and clinical duties.

Moreover, the system is highly rated for the educational environment for new rehabilitation staff. In the survey of novice staff, everyone evaluated that the system was helpful for learning and growth, and the median System Usability Scale (SUS) score of 90 points indicated excellent usability.[25] Free comments included opinions such as "can receive specialized and accurate feedback", "can receive feedback easily from anywhere", and "mental burden is reduced". Previous studies have also reported that the use of LLMs has a positive impact on improving learning skills, enhancing learning efficiency, and cognitive and emotional motivation.[11,23,26] Additionally, in medical education, its potential as a

powerful medical writing tool for generating summaries, proofreading, and providing medical insights has been noted.[27,28] Based on these findings, we believe that this system could become a powerful tool for improving the learning efficiency of novice rehabilitation staff.

Several challenges regarding the content of the meta-prompt and the use of the system have come to light, providing specific guidelines for future improvements. Firstly, the usage logs of the "correcting integration and interpretation" bot revealed that, at the initial report submission stage, the bot had judged all checkpoints as "passed" for most novice staff. This ceiling effect means that the bot over-estimated the novice staff skills. Secondly, several instructors expressed concerns that dependence on the system could lead to a decline in writing skills and clinical reasoning abilities. Previous studies have also highlighted risks in utilizing LLMs in educational settings, such as the loss of critical thinking, problem-solving abilities, and communication skills, as well as a lack of consistency in the output content and insufficient specialized knowledge.[23,28,29] Considering these challenges, in developing this system, we determined that providing individualized and specialized feedback would be difficult, and thus focused primarily on feedback related to "overall structure and writing style". In light of these findings and limitations, to provide more specialized and effective feedback in the future, it is necessary to improve and verify the meta-prompt and other newer LLMs. Users of LLMs for educational purposes need to be aware of their advantages and limitations.

Limitation

This study has several limitations. Firstly, with only five participants and no control group, it was difficult to establish a clear causal relationship between the intervention and the results. Given the nature of this study, conducting a rigorous randomized controlled trial (RCT) is currently challenging. It is necessary to first evaluate a system with improved prompts using a quasi-experimental method, and then further refine the system based on those results. Once the system is sufficiently established, it would be desirable to consider implementing an RCT.

Secondly, as this was a single-center study, the generalizability of the results is uncertain. Future research should involve multi-center collaborative studies to assess generalizability across different institutions.

Thirdly, the research outcomes were based on subjective data from questionnaires, which may have been influenced by respondent bias. Therefore, it is important to collect and analyze quantitative data, such as the number of times instructors provided guidance and the duration of their guidance. Additionally, having independent third reviewers

evaluate the reports before and after LLM feedback could yield more objective results.

Conclusion

The feedback system for case reports using LLM has the potential to reduce the burden on educational instructors and contribute to providing an efficient educational environment for novice rehabilitation staff. In the future, it will be necessary to create a revised version of the system and conduct further evaluation.

Acknowledgments

For the English proofreading of this manuscript, we utilized Claude 3.5 Sonnet. All outputs were subsequently reviewed and editted by the authors to ensure accuracy and appropriateness.

We would like to thank our rehabilitation staff for their help with the study.

Authors' contributions

Yuuto Tonouchi (first author): Writing – original draft, conceptualization, methodology, project administration

Shunsuke Nakai: Writing - review & editing, investigation, validation,

Kayo Murakami: Writing - review & editing, investigation, validation,

Yuki Kataoka: Writing - review & editing, conceptualization, methodology, resources, funding acquisition, supervision

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis, and interpretation, or in all these areas; took part in drafting, revising, or critically reviewing the article; gave final approval of the version to be published; agreed on the journal to which the article has been submitted; and agreed to be accountable for all aspects of the work.

Statements and declarations

Ethical considerations

The study protocol was approved by the Ethics Committee of Hospital A, and the study was conducted after obtaining individual consent from the participating staff members (ID: 2024-0502). This study was also pre-registered in the UMIN Clinical Trials Registry (UMIN-CTR) (Trial ID: UMIN000053315).

Consent to participate

Informed consent was obtained from all participants in writing.

Consent for publication

Not applicable

Declaration of conflicting interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and publication of this article

Funding statement

The application programming interface fee was supported by a research fund of Scientific Research Works Peer Support Group (SRWS-PSG) provided to YK. The funder played no role in the study design, data collection and analysis, publication decisions, or manuscript preparation.

Data Availability

The data supporting the findings of this study, including the developed meta-prompts and code, are available on GitHub (https://github.com/youkiti/report-feedback). Other data are available upon request to the corresponding authors.

Reference

1. Oermann MH, Garvin MF. Stresses and challenges for new graduates in hospitals. Nurse Educ Today 2002; 22: 225–230.

2. Parker V, Giles M, Lantry G, et al. New graduate nurses' experiences in their first year of practice. Nurse Educ Today 2014; 34: 150–156.

3. Karaman S. Nurses' perceptions of online continuing education. BMC Med Educ 2011; 11: 86.

4. Shih YS, Lee TT, Liu CY, et al. Evaluation of an online orientation program for new healthcare employees. Comput Inform Nurs 2013; 31: 343–350.

5. Ministry of Health, Labour and Welfare. Guidelines for Training New Nursing Staff [Revised Edition], https://www.mhlw.go.jp/file/06-Seisakujouhou-10800000-Iseikyoku/0000049466_1.pdf (2014, accessed 6 June 2024).

6. Japanese Nursing Association. Learning Support Book for Nurses, https://www.nurse.or.jp/nursing/assets/learning/support-learning-guide-all.pdf (2023, accessed 6 June 2024).

7. Japanese Physical Therapy Association. Guidelines for Training New Physical Therapist Staff (First Edition), https://www.japanpt.or.jp/assets/pdf/pt/lifelonglearning/introeduprogram/education_training/training_guidelines_201111.pdf (2020, accessed 6 June 2024).

8. Shiota S, Goto N, Kanayama A, et al. Current Status and Challenges of Support Environments for New Graduate Occupational Therapists in Japanese Hospitals. A Mixed Method Study. Occup Ther Int 2022; 2022: 2159828.

9. Japanese Association of Occupational Therapists. Guidelines for Occupational Therapy Clinical Training (2018) / Handbook for Occupational Therapy Clinical Training (2022), https://www.jaot.or.jp/files/shishin2018.tebiki2022.2.pdf (2022, accessed 6 June 2024).

10. Suzuki Y, Horimoto Y. Survey of the Actual Conditiond of Newcomer in Medical Facilities. Rigakuryoho Kagaku 2022; 37: 375–382.

11. Dai W, Lin J, Jin F, et al. Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, Orem, UT, USA, 10–13 July 2023, pp. 323–325.

12. Liang W, Zhang Y, Cao H, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI 2024; AIoa2400196.

13. Nikkei XTECH: AI to Review Contracts, Competition in Legal Tech Intensifies with "Seal of Approval" from the Ministry of Justice, https://xtech.nikkei.com/atcl/nxt/column/18/00001/08386/ (2023, accessed 6 June 2024)

14. Humza N, Khan AU, Qiu S, et al. "A Comprehensive Overview of Large Language Models." arXiv:2307.06435v9. Epub ahead of print 9 Apr 2024. DOI: https://doi.org/10.48550/arXiv.2307.06435.

15. Ivankova NV, Creswell JW, Stick SL. Using Mixed-Methods Sequential Explanatory Design: From Theory to Practice. Field Methods 2006; 18: 3–20.

16. Zhang Y, Yuan Y, Yao AC. Meta Prompting for AI Systems. arXiv:2311.11482v5. Epub ahead of print 2 Apr 2024. DOI: https://doi.org/10.48550/arXiv.2311.11482.

17. von Elm E, Altman DG, Egger M, et al.; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)statement: guidelines for reporting observational studies. J Clin Epidemiol 2008; 61: 344–349.

18. Anthropic: User Guides, https://docs.anthropic.com/en/docs/welcome (2024, accessed 6 June 2024)

19. Slack: Where work happens, https://slack.com/ (2017, accessed 11 June 2024)

20. Japanese Law Translation. Act on the Protection of Personal Information, Act on the Protection of Personal Information - Japanese/English - Japanese Law Translation (2003, accessed 8 August 2024)

21. Anthropic: Prompt Library Prose polisher, https://docs.anthropic.com/en/prompt-library/prose-polisher (2024, accessed 6 June 2024)

22. Younas A, Subramanian KP, Al-Haziazi M, et al. A Review on Implementation of Artificial Intelligence in Education. International Journal of Research and Innovation in Social Science 2023; 7: 1092–1100.

23. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. J Educ Eval Health Prof 2024; 21: 6.

24. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences 2023; 103: 102274.

25. Lewis, J. R. The System Usability Scale: Past, Present, and Future. International Journal of Human–Computer Interaction 2018; 34: 577–590.

26. Meyer J, Jansen T, Schiller R, et al. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Computers and Education: Artificial Intelligence 2024; 6: 100199.

27. Ho WLJ, Koussayer B, Sujka J. ChatGPT: Friend or foe in medical writing? An example of how ChatGPT can be utilized in writing case reports. Surgery in Practice and Science 2023; 14: 100185.

28. Abd-alrazaq A, AlSaad R, Alhuwail D, et al. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. JMIR Med Educ 2023; 9: e48291.

29. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ 2024; 17: 926-931.