AI's Accuracy in Extracting Learning Experiences from Clinical Practice Logs: An Observational Study

Takeshi Kondo, Hiroshi Nishigori

Submitted to: JMIR Medical Education
on: November 17, 2024

Table of Contents

AI’s Accuracy in Extracting Learning Experiences from Clinical Practice Logs: An Observational Study

Takeshi Kondo[1] MD, MHPE; Hiroshi Nishigori[1] MD, PhD

[1]Nagoya University Nagoya JP

Corresponding Author:
Takeshi Kondo MD, MHPE
Nagoya University
65, Tsurumai-cho, Showa-ku
Nagoya
JP

Abstract

Background: Improving the quality of education in clinical settings requires an understanding of learners’ experiences and learning processes. However, this is a significant burden on learners and educators. If learners’ learning records could be automatically analyzed and experiences visualized, it would enable real-time tracking of their progress. Large language models (LLMs) may be useful for this purpose, although their accuracy has not been sufficiently studied.

Objective: This study aimed to explore the accuracy of predicting the actual clinical experiences of medical students from their learning log data during clinical clerkship using LLMs.

Methods: This study was conducted at the Nagoya University School of Medicine. Learning log data from medical students participating in a clinical clerkship from April 22, 2024, to May 24, 2024, were used. The Model Core Curriculum (MCC) for Medical Education was employed as a template to extract experiences. OpenAI’s ChatGPT was selected for this task after a comparison with other LLMs. Prompts were created using the learning log data and provided to ChatGPT to extract experiences, which were then listed. A web application using GPT-4-turbo was developed to automate this process. The accuracy of the extracted experiences was evaluated by comparing them with the corrected lists provided by the students.

Results: Twenty out of thirty-three 6th-year medical students participated in this study, resulting in 40 datasets. The Jaccard Index was 0.59, indicating moderate agreement. Sensitivity and specificity were 62.67% and 99.37%, respectively. The results suggest that GPT-4-turbo accurately identifies many of the actual experiences but misses some because of insufficient detail or a lack of student records.

Conclusions: This study demonstrated that LLMs, such as GPT-4-turbo, can predict clinical experiences from learning logs with high specificity but moderate sensitivity. Future improvements in AI models and the combination of learning logs with other data sources, such as electronic medical records, may enhance the accuracy. Utilizing naturally accumulated data for assessment could reduce the burden on learners and educators while improving the quality of educational assessments in medical education.

(JMIR Preprints 17/11/2024:68697)
DOI: https://doi.org/10.2196/preprints.68697

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

AI's Accuracy in Extracting Learning Experiences from Clinical Practice Logs: An Observational Study

Takeshi Kondo, MD, MHPE and Hiroshi Nishigori, MD, PhD

Corresponding author:
Takeshi Kondo, MD, MHPE
Department of General Medicine/Family & Community Medicine, Nagoya University Graduate School of Medicine, Nagoya, Japan
65, Tsurumai-cho, Showa-ku,
Nagoya-city, Aichi,
Japan
E-mail: ncukondo@gmail.com;
Tel: +81-52-741-2111
Fax: +81-52-741-2111
ORCID ID: https://orcid.org/0000-0002-3307-671X

Abstract

Background: Improving the quality of education in clinical settings requires an understanding of learners' experiences and learning processes. However, this is a significant burden on learners and educators. If learners' learning records could be automatically analyzed and experiences visualized, it would enable real-time tracking of their progress. Large language models (LLMs) may be useful for this purpose, although their accuracy has not been sufficiently studied.
Objective: This study aimed to explore the accuracy of predicting the actual clinical experiences of medical students from their learning log data during clinical clerkship using LLMs.
Methods: This study was conducted at the Nagoya University School of Medicine. Learning log data from medical students participating in a clinical clerkship from April 22, 2024, to May 24, 2024, were used. The Model Core Curriculum (MCC) for Medical Education was employed as a template to extract experiences. OpenAI's ChatGPT was selected for this task after a comparison with other LLMs. Prompts were created using the learning log data and provided to ChatGPT to extract experiences, which were then listed. A web application using GPT-4-turbo was developed to automate this process. The accuracy of the extracted experiences was evaluated by comparing them with the corrected lists provided by the students.
Results: Twenty out of thirty-three 6th-year medical students participated in this study, resulting in 40 datasets. The Jaccard Index was 0.59, indicating moderate agreement. Sensitivity and specificity were 62.67% and 99.37%, respectively. The results suggest that GPT-4-turbo accurately identifies many of the actual experiences but misses some because of insufficient detail or a lack of student records.
Conclusions: This study demonstrated that LLMs, such as GPT-4-turbo, can predict clinical experiences from learning logs with high specificity but moderate sensitivity. Future improvements in AI models and the combination of learning logs with other data sources, such as electronic medical records, may enhance the accuracy. Utilizing naturally accumulated data for assessment could reduce the burden on learners and educators while improving the quality of educational assessments in medical education.

Keywords: Large Language Models; ChatGPT; workplace-based assessment; big data-based assessment

Introduction

To improve the quality of education in clinical settings, it is important to understand what learners experience and how they learn [1,2]. Various workplace-based assessment tools have been developed and utilized to enable educators to track learners' progress and provide feedback [3]. However, the rigorous management of learners' progress requires frequent observation of learners, frequent evaluations, and feedback from educators. This can impose a high burden on both learners and educators, potentially hindering learning [4,5]. Thus, the challenge is accurately monitoring learning in clinical settings without burdening learners or educators.

Learners in clinical settings often document their learning and practice experiences. If these records can be analyzed to understand learners' contexts, monitoring their learning without imposing additional burdens may be possible. One such record kept by learners during clinical clerkship is a logbook. The logbook documents the cases encountered, procedures performed, and learner's reflections. It serves as a tool for prompting student reflections and facilitating feedback and dialogue between educators and learners [6-8]. Evaluating these records against curriculum competencies and goals without adding an extra burden on learners can help monitor their progress. However, educators may have to manually match and analyze these records, which may be a significant burden.

AI-assisted text extraction and standard matching could be useful in this context. Previous studies have successfully used natural language processing, a branch of AI, to analyze supervisory feedback comments and predict student performance against competency standards [9]. AI models that integrate multiple information sources to represent student performance have also been developed [10]. Among AI technologies, large language models (LLMs) have gained attention in medical education because of their extensive pretraining on large datasets, allowing them to handle various situations, including multilingual support, with minimal adjustment [11]. Research using ChatGPT, an LLM, has shown that it can apply codes to interview texts using a codebook, suggesting its potential for extracting competency-based evaluations from student descriptions [12]. These studies indicate that LLMs can potentially extract competency-related experiences from student descriptions. However, owing to a lack of such research, aggregation accuracy remains uncertain. Determining the extent to which LLMs can aggregate items related to curriculum goals from learner descriptions may open opportunities to leverage LLMs to monitor learner progress and enhance education quality.

Therefore, this study focuses on undergraduate clinical clerkships in Japan to investigate the accuracy with which LLMs can aggregate goals from records kept for learning. In Japanese pre-graduate education, the Medical Education Model Core Curriculum (MCC) [13] was established to define two-thirds of the undergraduate curriculum and is used as a guideline for pre-graduate medical education. The MCC outlines the experiences medical students should have by the time they graduate, focusing primarily on clinical clerkships [13]. Therefore, our research question is: how accurately can an LLM predict experiences related to the goals defined by the MCC from the records students keep for learning during clinical clerkships?

Methods

Context

This study was conducted as part of the clinical participatory clinical clerkship at the Nagoya University School of Medicine, a program designed to provide medical students with practical experience in clinical settings. During the final year of medical school (6th year), students participate in this program for four weeks, recording their daily experiences and learning activities. A trial to transform these records into an electronic portfolio began in 2024. This study was part of this trial.

Dataset

This study used learning log data from 6th-year medical students to extract their experiences related to core curriculum goals. Learning log data consisted of daily records of experiences and learning activities entered by medical students into an electronic portfolio during a clinical clerkship from April 22, 2024, to May 24, 2024. The data were treated as weekly datasets.

Extraction of Experiences

The template for extracting experiences from the dataset was the MCC for medical education [13]. This study used a table of symptoms, examinations, and procedures that medical students are expected to experience during their clinical clerkship at Nagoya University as the template for experience extraction (Table 1).

Table 1. Symptoms, examinations, and procedures that medical students are expected to experience

Categories	Item
Symptoms	
	Fever
	General malaise
	Anorexia
	Weight loss
	Weight gain
	Altered mental status
	Syncope
	Seizure
	Vertigo and dizziness
	Edema
	Rash
	Cough and sputum production
	Blood in sputum and hemoptysis
	Dyspnea
	Chest pain
	Palpitations

	Dysphagia
	Abdominal pain
	Nausea and vomiting
	Hematemesis
	Melena
	Constipation
	Diarrhea
	Jaundice
	Abdominal distention and abdominal mass
	Lymphadenopathy
	Abnormal urine output/urination
	Hematuria
	Menstrual abnormality
	Anxiety/depression
	Cognitive dysfunction
	Headache
	Skeletal muscle paralysis/muscle weakness
	Gait disturbance
	Sensory disturbance
	Back pain
	Arthralgia/joint swelling
Examinations	
	Full blood count
	Blood biochemistry
	Coagulation/fibrinolysis
	Immunoserology tests
	Urinalysis
	Stool (fecal) examination
	Blood typing (ABO, RhD), blood compatibility test (cross-matching), atypical antibody screening
	Arterial blood gas analysis
	Pregnancy test
	Microbiological tests (bacterial smear, culture, identification, antibiotic sensitivity test)
	Cerebrospinal fluid
	Pleural fluid analysis
	Peritoneal fluid analysis
	Histopathology and cytology (including intraoperative rapid diagnosis)
	Genetic testing and chromosome analysis

	ECG
	Lung function tests
	Endocrine and metabolic function tests
	Electroencephalography
	Ultrasound
	X-ray
	CT
	MRI
	Nuclear medicine examination
	Endoscopy
Procedures	
	Position change, transfer
	Skin antisepsis
	Application of topical medications
	Airway suction
	Nebulizer
	Venous blood sampling
	Peripheral venous catheterization
	Insertion and extraction of nasogastric tube
	Insertion and extraction of urinary catheter
	Intradermal injection
	Subcutaneous injection
	Intramuscular injection
	Intravenous injection
	Urinalysis (including pregnancy test)
	Microbiological testing (including gram staining)
	Recording of a 12-lead ECG
	Rapid bedside ultrasound (including FAST) for clinical decision-making
	Rapid antigen/pathogen testing
	Blood glucose test
	Aseptic technique
	Surgical hand washing
	Gowning techniques in the operating room
	Basic sutures and suture removal

OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude were considered for the LLMs used in experience extraction. ChatGPT by OpenAI, which consistently provided output in a uniform format, was adopted after comparing these tools with randomly selected student records and the

extraction template.

LLMs, including ChatGPT, receive text data as input and generate subsequent text based on this data. Therefore, the prompt given to the LLM is crucial. In this study, prompts were created using medical students' learning log data, which were provided to ChatGPT to extract their experiences from the logs. Experiences were extracted based on a table of symptoms, tests, and procedures that students were expected to experience, with ChatGPT outputting a list of symptoms, tests, and procedures inferred from the text data. To automate this process, a web application using GPT-4-turbo was developed, which allows medical students to input learning log data and receive the extracted experiences as a list output from GPT-4-turbo (gpt-4-0125-preview).

Evaluation of Extracted Experiences

The extracted experience goals were presented to the medical students via e-mail. Students were asked to compare the list with their actual experiences, including those not recorded in their reflections, and to submit a corrected list. The corrected lists were compared with the original learning log data to evaluate the accuracy of the extracted experiences.

Data Analysis

The accuracy of the extracted experience goals was evaluated using R software (version 4.1.2). The agreement rate between the extracted and corrected experience goals was calculated, and the accuracy of the extracted experience goals was assessed based on this agreement rate.

Ethical Considerations

This study was approved by the Ethics Committee of Nagoya University Graduate School of Medicine (Approval Number: 2023-0451 31742). All participants were informed about the study's purpose, methods, risks, and benefits and were allowed to opt-out.

Results

During the clinical participation-based clerkship at Nagoya University Hospital from April 22, 2024, to May 24, 2024, 20 out of 33 sixth-year students who made entries in the e-portfolio participated in the study, yielding 40 data points. The predicted experiences and actual experiences are shown in Table 2.

Table 2. Predicted experience items and actual experience items

Index	Predicted items	Actual items	Match	Predicted	Actual
1	Skeletal muscle paralysis/muscle weakness, Gait disturbance, Sensory disturbance	Skeletal muscle paralysis/muscle weakness, Gait disturbance, Sensory disturbance	3	3	3
2	Endocrine and metabolic function tests	Endocrine and metabolic function tests	1	1	1

3	Fever	Fever	1	1	1
4	Basic sutures and suture removal	Basic sutures and suture removal	1	1	1
5	Seizure, Electroencephalography, MRI	Aseptic technique, Electroencephalography, MRI, Weight gain, Seizure	3	3	5
6	Skeletal muscle paralysis/muscle weakness, Gait disturbance, Sensory disturbance	Skeletal muscle paralysis/muscle weakness, Gait disturbance, Sensory disturbance	3	3	3
7	Anorexia, Abdominal distention, and abdominal mass, Ultrasound	Palpitations, Skeletal muscle paralysis/muscle weakness	0	3	2
8	Venous blood sampling	Venous blood sampling	1	1	1
9	Rapid bedside ultrasound (including FAST) for clinical decision-making, Ultrasound	Skin antisepsis, Rapid bedside ultrasound (including FAST) for clinical decision-making, Aseptic technique, Surgical hand washing, Gowning techniques in the operating room, Basic sutures and suture removal, Ultrasound, Fever, Diarrhea	2	2	9
10	Basic sutures and suture removal	Basic sutures and suture removal	1	1	1
11	Basic sutures and suture removal	Surgical hand washing, Gowning techniques in the operating room, Basic sutures and suture removal, Full blood count, Blood biochemistry, Coagulation/fibrinolysis, Histopathology and cytology (including intraoperative rapid diagnosis), X-ray, CT, MRI, General malaise, Weight loss	1	1	12
12	Venous blood	Venous blood	1	2	1

	sampling, Pregnancy test	sampling			
13	Surgical hand washing, Gowning techniques in the operating room, Basic sutures and suture removal	Surgical hand washing, Gowning techniques in the operating room, Basic sutures and suture removal	3	3	3
14	Pregnancy test, Basic sutures and suture removal	Position change, transfer, Insertion and extraction of a urinary catheter, Surgical hand washing, Gowning techniques in the operating room, Basic sutures and suture removal, Histopathology and cytology (including intraoperative rapid diagnosis), MRI, Abdominal distention and abdominal mass	1	2	8
15	Surgical hand washing	Surgical hand washing, Gowning techniques in the operating room, Basic sutures and suture removal, Full blood count, Blood biochemistry, Histopathology and cytology (including intraoperative rapid diagnosis), Ultrasound, X-ray	1	1	8
16	Microbiological tests (bacterial smear, culture, identification, antibiotic sensitivity test), Nuclear medicine examination, General malaise, Cough and sputum production, Dyspnea, Abdominal pain, Nausea and vomiting, Abnormal urine output/urination	General malaise, Edema	1	8	2

17	Surgical hand washing	Aseptic technique, Surgical hand washing, Gowning techniques in the operating room, Basic sutures and suture removal, Cough and sputum production	1	1	5
18	Fever, Urinalysis, Microbiological tests (bacterial smear, culture, identification, antibiotic sensitivity test), Nausea and vomiting, Hematuria	Full blood count, Blood biochemistry, Immunoserology tests, Urinalysis, Microbiological tests (bacterial smear, culture, identification, antibiotic sensitivity test), Edema, Palpitations, Hematuria, Back pain	3	5	9
19	Blood glucose test, Endocrine and metabolic function tests	Blood glucose test, Endocrine and metabolic function tests	2	2	2
20	Cognitive dysfunction	Cognitive dysfunction	1	1	1
21	Chest pain	Chest pain	1	1	1
22	Surgical hand washing, Basic sutures and suture removal	Aseptic technique, Surgical hand washing, Gowning techniques in the operating room, Basic sutures and suture removal	2	2	4
23	Cognitive dysfunction, Abnormal urine output/urination, Urinalysis	Cognitive dysfunction, Abnormal urine output/urination	2	3	2
24	Dyspnea	Dyspnea	1	1	1
25	Gowning techniques in the operating room	Position change, transfer, Skin antisepsis, Aseptic technique, Surgical hand washing, Gowning techniques in the operating room, Basic sutures and suture removal	1	1	6
26	Full blood count, Blood biochemistry	Full blood count, Blood biochemistry, Immunoserology tests,	2	2	4

		Edema			
27	CT, MRI, X-ray	Position change, transfer, Full blood count, Arterial blood gas analysis, Ultrasound, X-ray, CT, MRI, Skeletal muscle paralysis/muscle weakness, Gait disturbance, Back pain	3	3	10
28	Endocrine and metabolic function tests	Endocrine and metabolic function tests	1	1	1
29	Basic sutures and suture removal	Position change, transfer, Surgical hand washing, Gowning techniques in the operating room, Basic sutures and suture removal	1	1	4
30	Weight loss, Skeletal muscle paralysis/muscle weakness	Blood glucose test, Weight loss, Skeletal muscle paralysis/muscle weakness	2	2	3
31	Ultrasound, Endoscopy	Ultrasound, Endoscopy	2	2	2
32	Basic sutures and suture removal	Skin antisepsis, Aseptic technique, Surgical hand washing, Gowning techniques in the operating room, Basic sutures and suture removal, Full blood count, Blood biochemistry, Coagulation/fibrinolysis, Immunoserology tests, Histopathology and cytology (including intraoperative rapid diagnosis), Ultrasound, X-ray, Headache	1	1	13
33	Skin antisepsis, Position change, transfer	Skin antisepsis, Position change, transfer	2	2	2

34	Back pain	Weight loss, Cognitive dysfunction, Skeletal muscle paralysis/muscle weakness, Sensory disturbance, Back pain	1	1	5
35	Arterial blood gas analysis, Peripheral venous catheterization, Insertion and extraction of a nasogastric tube, Insertion and extraction of a urinary catheter, Aseptic technique, Surgical hand washing, Gowning techniques in the operating room, Basic sutures and suture removal	Peripheral venous catheterization, Aseptic technique, Full blood count, Blood biochemistry, Coagulation/fibrinolysis, Arterial blood gas analysis, Pleural fluid analysis, Ultrasound, X-ray, CT, Endoscopy	3	8	11
36	Weight gain, Endocrine and metabolic function tests, Blood glucose test	Blood glucose test, Full blood count, Blood biochemistry, Urinalysis, Stool (fecal) examination, Endocrine and metabolic function tests, Ultrasound, CT, Weight gain	3	3	9
37	Endoscopy	Endoscopy	1	1	1
38	X-ray	X-ray, Cough and sputum production	1	1	2
39	Abdominal pain	id not found	0	1	1
40	Skin antisepsis	Skin antisepsis	1	1	1

The predicted items were experience items extracted using GPT-4-turbo from the students' practice records. The actual items were those that the students checked as experiences they had during that period. "Match" refers to the number of matches, "predicted" refers to the number of experience items extracted by GPT-4-turbo, and "actual" refers to the number of items the students checked as experiences they actually had. The students' practice records used by GPT-4-turbo to extract experiences are presented in Appendix 1, with the "Index" column in Table 2 corresponding to the "Index" column in Appendix 1. The Jaccard Index was 0.59, indicating moderate agreement. Sensitivity and specificity were 62.67% and 99.37%, respectively.

Discussion

In this study, we analyzed the records kept by medical students during their clinical clerkship for

learning purposes, using the GPT-4-turbo to predict the clinical procedures they experienced. The experiences extracted by the GPT-4-turbo were evaluated for accuracy after being revised by the medical students. The high specificity of the predictions made by GPT-4-turbo suggests that the extracted experiences likely mirror what the students actually encountered.

Previous studies have explored the use of logbooks to monitor learner situations. Attempts have been made to monitor skills and experiences using logbooks [14], track the progress of Entrustable Professional Activities (EPAs) [8], and count the cases encountered [7]. However, the "logbooks" used in these studies are lists of cases experienced or evaluations rather than detailed descriptions of experiences [7,8]. This format is more useful for evaluation purposes rather than for recording learning, which ultimately adds to the burden on learners. Our study suggests that analyzing reflections purely recorded for learning purposes can also extract experiences, offering a technique that monitors learning situations while reducing the burden on learners and educators. However, the low sensitivity suggests that some of the students' actual experiences may not have been captured by GPT-4-turbo's analysis.

There were several patterns of experiences that GPT-4-turbo could not extract despite being actually experienced by the students. The first pattern involved experiences that could have been predicted from the records but were missed by GPT-4-turbo. For example, one student documented an experience with a hereditary amyotrophic lateral sclerosis (ALS) case, but the GPT-4-turbo failed to show that the student recorded muscle weakness, a symptom of ALS. Another instance involved a student learning about palpitations as a symptom of Takayasu's arteritis; however, GPT-4-turbo did not extract palpitations from the record. The second pattern involved experiences in which predictions were difficult due to insufficient documentation. Many students recorded observing surgeries, but it was unclear whether they assisted in the surgery or merely observed it from outside, making it difficult for GPT-4-turbo to extract actions such as surgical handwashing and gown techniques. The third pattern involved experiences that were not documented by the students, making predictions impossible. One student noted observing a surgery but actually performed suturing, which was not recorded. Another student documented examining a diabetic patient but did not mention performing CT or ultrasound examinations, which were indeed performed. Improvements in AI model performance are expected to address the first pattern. However, different approaches are required for the more prevalent second and third patterns. Encouraging students to write more detailed reflections and combining other data, such as electronic health records (EHRs) written by the students, might be effective in this regard. Feeding both learning logs and EHR descriptions into the GPT-4-turbo could enhance the accuracy of experience extraction.

Nevertheless, caution should be exercised when using naturally accumulated data for evaluation. The first concern is the protection of personal information. Clinical practice and learning descriptions may contain personal information about patients and students, which must be appropriately protected. The second concern is the quality of evaluation. Naturally accumulated data differ from the data intentionally collected for evaluation, necessitating an examination of whether such data analysis provides reliable and valid evaluation data before actual use. The third concern is transparency. When utilizing naturally accumulated data for evaluation, it is essential to thoroughly explain the data collection and analysis methods to learners and educators. This transparency will help them understand how evaluation results are obtained, fostering trust and the effective use of the results by learners and educators.

Limitations

This study has several limitations. First, the study used learning log data from clinical participation-based clinical clerkships at a single university; therefore, its generalizability to learning log data from other universities or clinical clerkships is not guaranteed. Additionally, while the accuracy of the extracted experience content was evaluated using learning log data recorded by medical students and asking them to make corrections, the quality and quantity of the learning log data recorded by the students could affect the accuracy of the extracted experience content. Large-scale collaborative studies across multiple institutions are needed to ensure broader generalizability. Furthermore, this study used a list of symptoms, examinations, and procedures in the MCC as a template for extracting experience content; however, the results of using other templates were not examined. Future research is needed to assess the performance using other evaluation criteria. In this study, the accuracy of the extracted experience content was evaluated by using learning log data recorded by medical students and asking them to make corrections, but no strict criteria were set for what constitutes “experience” when students made corrections. Although the current method seems to yield data close to the experiences perceived by the students, more rigorous verification requires stricter criteria regarding what students consider as “experience.”

Conclusions

In this study, records kept by medical students for learning during clinical clerkships were analyzed using the GPT-4-turbo to predict experienced clinical activities. The high specificity of the GPT-4-turbo predictions suggests that the extracted experiences are likely what students actually experienced. However, the low sensitivity indicates that some actual student experiences were not captured by the GPT-4-turbo analysis. Future utilization of accumulated data for learning or practice may enable detailed assessments while avoiding excessive burdens on learners and educators.

Acknowledgments

TK was responsible for the study planning, data collection, analysis, and manuscript writing. NH collaborated with TK on the study planning and provided supervision and advice on data analysis and manuscript writing. ChatGPT was used in part to create an initial English translation of the Japanese version of the manuscript. This work was supported by JSPS KAKENHI Grant Number 23K27816.

Conflicts of Interest

The authors have no conflicts of interest to declare.

Abbreviations

MCC: the Medical Education Model Core Curriculum
LLMs: Large Language Models
SLMs: Small Language Models

Multimedia Appendix 1

The GitHub repository includes an experience extraction API and R code to analyze the and the data itself. https://github.com/ncukondo/extract-experiences-by-llm

References

1. AlHaqwi AI, Taha WS. Promoting excellence in teaching and learning in clinical education. J Taibah Univ Med Sci 2015;10(1):97-101
2. Vanka A, Hovaguimian A. Teaching strategies for the clinical environment. Clin Teach 2019; 16(6):570-574.doi:10.1111/tct.12928
3. Liu C. An introduction to workplace-based assessments. Gastroenterol Hepatol Bed Bench 2012;5(1):24-28.
4. Ott MC, Pack R, Cristancho S, Chin M, Van Koughnett JA, Ott M. “The most crushing thing”: Understanding resident assessment burden in a competency-based curriculum. J Grad Med Educ 2022;14(5):583-592. doi:10.4300/JGME-D-22-00050.1
5. Szulewski A, Braund H, Dagnone DJ, et al. The assessment burden in competency-based medical education: How programs are adapting. Acad Med 2023;98(11):1261-1267. doi:10.1097/ACM.0000000000005305
6. Alotaibi HM, Alharithy R, Alotaibi HM. Importance of the reflective logbook in improving the residents’ perception of reflective learning in the dermatology residency program in Saudi Arabia: findings from a cross-sectional study. BMC Med Educ 2022;22(1):862. doi:10.1186/s12909-022-03948-w
7. Alabbad J, Abdul Raheem F, Almusaileem A, Almusaileem S, Alsaddah S, Almubarak A. Medical students’ logbook case loads do not predict final exam scores in surgery clerkship. Adv Med Educ Pract 2018;9:259-265. doi:10.2147/AMEP.S160514
8. Berberat PO, Rotthoff T, Baerwald C, et al. Entrustable Professional Activities in final year undergraduate medical training - advancement of the final year training logbook in Germany. GMS J Med Educ 2019;36(6):Doc70. doi:10.3205/zma001278
9. Gin BC, Ten Cate O, O’Sullivan PS, Hauer KE, Boscardin C. Exploring how feedback reflects entrustment decisions using artificial intelligence. Med Educ 2022;56(3):303-311. doi:10.1111/medu.14696
10. Millán E, Loboda T, Pérez-de-la-Cruz JL. Bayesian networks for student model engineering. Comput Educ 2010;55(4):1663-1683. doi:10.1016/j.compedu.2010.07.010
11. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel) 2023;11(6). doi:10.3390/healthcare11060887
12. Xiao Z, Yuan X, Liao QV, Abdelghani R, Oudeyer PY. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces. IUI ’23 Companion. Association for Computing Machinery 2023;75-78.
13. Medical Education Model Core Curriculum Expert Research Committee. The Model Core Curriculum for Medical Education (2022 Revision). Published 2022. http://jsme.umin.ac.jp/eng/core-curriculum.html
14. Levine RB, Kern DE, Wright SM. The impact of prompted narrative writing during internship on reflective practice: a qualitative study. Adv Health Sci Educ Theory Pract 2008; 13(5):723-733. doi.org:10.1007/s10459-007-9079-x