Original Paper

Evaluating Source-Based Large Language Models for Preclinical Dermatology Education: Comparative Study

Frank Je-Min Lin[1], BA; Sunghun Cho[2], MD

[1]F. Edward Hébert School of Medicine, Uniformed Services University of the Health Sciences, Bethesda, MD, United States
[2]Department of Dermatology, Uniformed Services University of the Health Sciences, Bethesda, MD, United States

Corresponding Author:

Frank Je-Min Lin, BA
F. Edward Hébert School of Medicine
Uniformed Services University of the Health Sciences
4301 Jones Bridge Road
Bethesda, MD 20814
United States
Phone: 1 2532733100
Email: frankjmlin@gmail.com

Abstract

Background: Large language models (LLMs) have gained increasing popularity in medical education, with evidence supporting their educational value when framed through the lens of cognitive load theory. Source-based LLMs, which explicitly ground responses in user-uploaded material via retrieval-augmented generation algorithms, may offer additional educational value by using student-developed materials to conceptualize new areas of learning within a familiar framework. This has applications for areas like medical education in dermatology, which could benefit from inclusive sources and enhanced education to alleviate health care gaps. However, no prior studies have examined whether the inclusion of student-authored notes alters the response characteristics of a source-based LLM when responding to medical questions.

Objective: This study aims to conduct an observational, comparative performance evaluation study assessing the accuracy, response reproducibility, and intermodel response similarity of freely available LLMs on text-only step 1 dermatology questions, and to explore whether providing extensive student-generated notes to a source-based LLM alters these performance characteristics.

Methods: In December 2024, 4 LLMs were evaluated: NotebookLM (NLM) with uploaded preclerkship study guides (NLM w/ Notes), NLM with an uploaded blank sheet of paper (NLM w/o Notes), ChatGPT-4o Mini, and Google Gemini 1.5 Flash. Each model completed 3 trials of 121 text-based United States Medical Licensing Examination (USMLE) step 1 dermatology questions from the AMBOSS question bank. They were evaluated for overall majority-consensus accuracy, accuracy by question difficulty, intertrial reproducibility, and agreement in answer choice selection between models. Differences were analyzed through a Cochran Q omnibus test and subsequent pairwise McNemar tests with Benjamini-Hochberg correction. Response reproducibility and intermodel agreement were analyzed through Fleiss κ statistics with 95% CI.

Results: ChatGPT-4o Mini achieved the highest overall majority-consensus accuracy (102/121, 84.3%). NLM w/ Notes demonstrated the highest intertrial reproducibility (Fleiss κ=0.927, 95% CI 0.875-0.978) and strong performance on lower-difficulty questions but comparatively reduced accuracy on higher-difficulty items. NLM w/o Notes exhibited significantly higher omission rates (38/363, 10.5% vs ≤7/363, 1.92% for other models) than other tested LLMs. Sensitivity analysis excluding omissions increased NLM w/o Notes' accuracy from 66.9% (81/121) to 77.8% (77/99), matching NLM w/ Notes' accuracy of 74.4% (90/121). Intermodel agreement was significantly higher between NLM w/ Notes and ChatGPT-4o Mini compared to NLM w/o Notes and Gemini 1.5 Flash.

Conclusions: Provision of student-generated notes substantially increased response reproducibility in a source-based LLM, likely reflecting consistent retrieval of similar source excerpts across trials. However, note-grounding appeared to constrain performance on higher-difficulty questions, suggesting a retrieval-augmented generation algorithm retrieval error when question stems excluded characteristic "keywords" present in lower-difficulty items. The results highlight potential challenges of a student-level, cognitive load theory–grounded educational LLM that must deal with notes not curated by experts, balance source use and internal reasoning, and meaningfully appraise uploaded sources to assess a student's individual learning gaps.

JMIR Form Res 2026;10:e88008; doi: 10.2196/88008

Keywords: NotebookLM; artificial intelligence; AI; AI in the classroom; large language model; LLM; source-based LLM; retrieval-augmented generation; cognitive load theory

Introduction

Large language models (LLMs) are artificial intelligence (AI) systems that have gained recent popularity in medical schools [1], with the potential to integrate into classrooms and enhance student learning outcomes through adaptive learning and simulation [2-4]. Indeed, recent neurologic research with electroencephalography monitoring has shown that AI in education is able to optimize a learner's cognitive load and increase learning efficiency through adaptive feedback and scaffolding, spaced repetition, and multimodal data integration [5]. The authors of this paper explicitly described the effective usage of AI in education through the lens of cognitive load theory (CLT)—a framework focused on optimizing the function of working memory to accumulate knowledge held in long-term memory. Notably underrepresented in educational literature, however, is the use of source-based LLMs—AI models that explicitly base their output on user-uploaded material.

Since 2021, source-based LLMs like NotebookLM (NLM) have used a computational technique called retrieval-augmented generation (RAG) that searches and retrieves pertinent words from a knowledge source to guide the statistical algorithms underlying their responses to queries [6]. For NLM specifically, users can upload up to 50 files, each with a maximum size of 200 MB; no queries can be given to the AI until at least 1 source is uploaded. On its home page, NLM claims to provide "clear citations," limit secondary chat data usage, and generate responses grounded in uploaded sources while using a Gemini Pro base [7]. The third claim hints at the potential niche for source-based LLMs using RAG within the CLT framework.

CLT research suggests that certain subjects (like math or grammar) may pose a high intrinsic cognitive load due to a learner needing to grasp the complex interactions of concepts with each other or having high "element interactivity." Certain instructional designs, such as providing worked examples, may decrease element interactivity and promote learning [8].

A classroom-integrated, source-based AI that draws directly from a student's notes, phrasings, and knowledge to contextualize new concepts in familiar ways of thought has the potential to reduce element interactivity by connecting new information to existing schemas. This may hypothetically lower cognitive load and working memory resource demands to a higher degree than previously characterized, allowing for increased learner efficiency.

Beyond learning frameworks, NLM potentially addresses issues in citations and transparency found in other LLMs. LLMs have encountered issues in citations, with only 71% to 77% of citations generated by ChatGPT-4 being confirmed to exist, and a rate of 50% to 90% of cited responses not being fully supported by their sources [9,10]. By citing user-inputted sources with direct links to specific words, NLM appears to alleviate some of these concerns. With regard to transparency, LLMs often do not disclose or properly document their training data. Such training data may contain copyrighted material, personal data, biased or harmful content, or even user-generated private inputs that may manifest in generated outputs [11]. NLM grounds its answers in uploaded material and promises not to use inputted data for training.

One possible niche for source-based LLMs like NLM is in dermatology medical education. The previously mentioned qualities that source-based LLMs hold may address some of the issues of underrepresentation in medical education and demographic bias. Current dermatology education remains inconsistent across medical curricula, despite the high prevalence of skin diseases in primary care and regional shortages of dermatologists [12,13]. To address this, some have suggested improving the baseline dermatology skills of medical students; however, dermatology education at medical schools remains highly variable [12].

There is also the added layer of existing demographic bias within dermatology, where the bias toward minimally melanated skin in lectures, textbooks, and LLMs is well documented within educational literature [13-15]. A May 2025 paper harnessing LLMs to generate clinical vignettes yielded poor ratings in the category of "demographic bias," due to a lack of incorporation of demographic diversity [16]. Source-based LLMs may not only increase the learning efficiency of core dermatology concepts in medical school through contextualization, as mentioned previously, but also, with the provision of diverse educational materials, may help alleviate some of the demographic bias displayed in other LLM responses.

Together, the educational gaps and representational biases mentioned highlight why tools that can be grounded in customizable, diverse source material are of particular interest in dermatology education. Source-based LLMs have demonstrated improvements in accuracy on subject-specific performance when provided with professionally created reference materials [17]. However, differences in response reproducibility, accuracy based on difficulty, or answer choice selection after source provision remain relatively unexplored. Furthermore, no prior studies have examined whether the inclusion of extensive student-authored notes affects accuracy or response reproducibility when answering standardized medical questions. This represents a gap in the literature, given that the CLT-derived educational value of a source-based LLM like NLM hinges on its capacity to function as an effective adjunct at the student level, even with expected flaws and errors in the provided sources.

Therefore, this formative study had two primary objectives: (1) to evaluate the accuracy, intertrial response

reproducibility, and intermodel response similarity of freely available LLMs on text-only step 1 dermatology questions and (2) to explore whether providing extensive student-generated notes alters these performance characteristics in a source-based model.

Methods

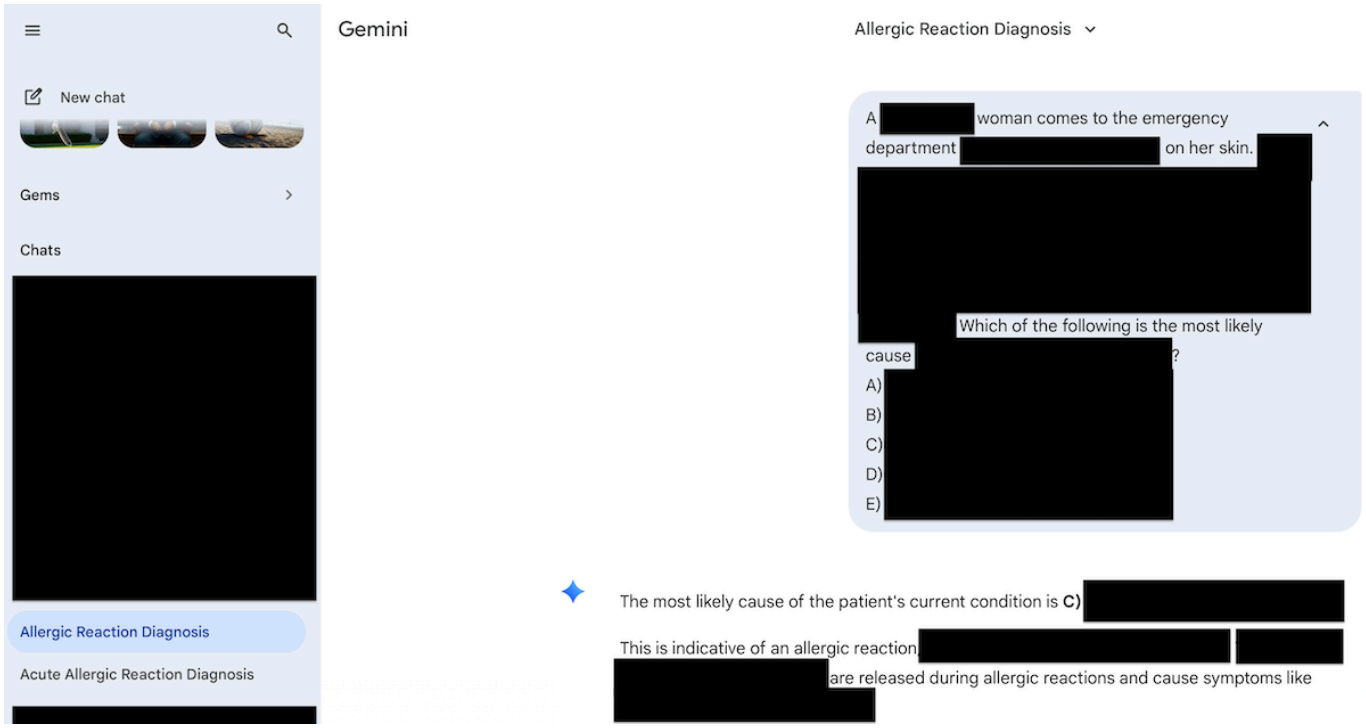
Overall Design

All available 182 QBank AMBOSS questions under the category of "USMLE Step 1 - Dermatology" were selected and generated in a "Study Session." Four questions were excluded because they contained tables that could not be copied and pasted, and 57 questions were excluded because they included pictures. In total, 121 questions were queried to each of the 4 LLMs tested. Three trials were conducted for each LLM, resulting in 363 total responses per model.

Overall accuracy and accuracy-by-difficulty for each model were determined by obtaining a question-level majority answer (for example, if an LLM answered with "D" in 2 or 3 of its trials, then the majority answer was "D") and comparing it against the answer key. If the majority answer matched the key, it was marked as "correct." Questions without a majority answer, or with an incorrect majority answer, were marked as "incorrect."

All questions were directly copied and pasted from AMBOSS into the LLM models without any additional queries (Figure 1).

Figure 1. Sample Gemini 1.5 Flash screenshot illustrating the query template used for each question in an observational, comparative performance evaluation of large language models (LLMs) answering United States Medical Licensing Examination (USMLE) step 1 dermatology questions conducted during December 2024. An AMBOSS question would be copied-and-pasted in its entirety as seen in the light gray box. Between each trial or attempt, the "new chat" button would be pressed in the top left to create a new instance.



A correct answer was defined as a response with either the correct answer choice or clearly demarcated answer text that exactly matched one of the answer choices. If the model did not select an answer, a new chat would be created, and the same question would be queried again until a definitive answer was selected, with a limit of three attempts. If the third attempt was completed and a definitive answer was still not selected, the trial result was recorded as an omission (X). For each individual trial, answers from the LLM would be tabulated as A, B, C, D, E, F, G, or X. The primary analysis used the question as the unit of analysis (N=121), with the majority-consensus response across the 3 trials checked against the answer key. This included questions that had trial-level omissions to reflect real-world model behavior. Further sensitivity analyses examining the majority response for each LLM, while excluding any question that had at least 1 trial-level omission from LLM-specific analysis, are displayed in Multimedia Appendix 1.

To minimize the influence of prior questions, the LLMs used were refreshed or prompted to create a new interaction for each query.

AMBOSS questions had preset difficulties. As described on the AMBOSS website, they were determined internally based on student performance on a scale of "1 Hammer" to "5 Hammer," with "1 Hammer" being the easiest 20% of questions, "2 Hammer" being between the easiest 20% and the upper 50% of all questions, "3 Hammer" being between the hardest 50% and 20% of all questions, "4 Hammer" being between the hardest 20% and 5% of all questions, and "5 Hammer" being the most difficult 5% of all questions [18].

The difficulty of each question was recorded and used for further statistical analysis.

This manuscript follows the STROBE (Strengthening the Reporting of Observational studies in Epidemiology) reporting guidelines where applicable [19] (Checklist 1).

LLM Details

The 4 LLMs used were: ChatGPT-4o Mini (OpenAI, December 2024 version), NLM w/ Notes, NLM w/o Notes, and Google Gemini 1.5 Flash (Google DeepMind, December 2024 version). ChatGPT-4o Mini was selected due to usage and query-limit restrictions on other ChatGPT variants. All 4 of these LLMs were accessed in their free, publicly available versions without subscription payment. The LLMs were used for experimentation only during the month of December 2024. At the time, no additional settings, including AI temperature, were available for modification at the user level. System-level prompts were similarly hidden from the user.

For the "NLM w/ Notes" LLM, 16 student-created text-only study guides for each preclerkship exam (totaling 216,117 words across 884 pages) were uploaded without modification. These guides, encompassing the entirety of the Uniformed Services University preclerkship curriculum, were developed collaboratively, without faculty involvement, by medical students at the Uniformed Services University over the past 5 years and stored on a shared, university-accessible Google Drive. They were created for educational use and shared voluntarily among students; no identifiable or proprietary content was included.

For the "NLM w/o Notes" LLM, a PDF of a blank 8.5"× 11" sheet (generated by exporting a blank Apple Pages sheet to PDF) was uploaded as the sole source document.

Statistical Analyses

An initial omnibus test was conducted to assess differences across all 4 LLMs for the overall and difficulty-level accuracy rates using multiple Cochran Q tests. Significant differences between LLMs were found for overall accuracy (Q_3=23.1215; P<.001), at the "2 Hammer" difficulty level (Q_3=16.9245; P<.001), and at the "3 Hammer" difficulty level (Q_3=11.3846; P=.009). The "4 Hammer" difficulty level did not show significant differences between LLMs (Q_3=2.7273; P=.44); "1 Hammer" and "5 Hammer" difficulty levels had identical results between all the LLMs.

Subsequent pairwise McNemar tests (without Yates correction) between LLMs (18 in total), α level of .05 after Benjamini–Hochberg (BH) correction, were limited to the previously mentioned groups that passed the Omnibus test. Cochran Q tests, McNemar tests, and BH corrections were calculated using R software (version 4.5.2; R Foundation for Statistical Computing). The *DescTools* R package was used for the Cochran *Q* test [20]. The full R code and the results of all 18 pairwise McNemar tests are available in Multimedia Appendix 2.

A Fleiss $\varkappa$ with a 95% CI of LLM-selected answer choices (A, B, C, D, E, F, G, or X) was calculated to assess the reliability of agreement between the 3 trials of each LLM, aiming to determine the reproducibility of their answers. Another Fleiss $\varkappa$ with a 95% CI was calculated to assess the reliability of agreement between all the trials of 2 LLMs (comparing 6 trials in total) to determine the pairwise LLM agreement between answer choices. Calculations were conducted using Google Sheets following the formula $\varkappa = (p_o - p_e)/(1 - p_e)$, where p_o represents observed agreement and p_e represents expected agreement. CIs around the sample mean were calculated at 95% CI using the formula $\bar{x} \pm 1.96 \times SE$, with SE approximated by Cochran (1960): $SE = \sqrt{(p_o(1- p_o)/(N(1- p_e)^2))}$ [21].

Ethical Considerations

Official permission to use the step 1 Qbank questions was granted by AMBOSS headquarters in Berlin, Germany. The experimental design (protocol DBS.2025.852; reference 980979) was reviewed by the Uniformed Services University's Human Research Protections Program Office and determined not to meet the criteria defining research per 32 CFR 219.102 and DoDI 3216.02. As such, the protocol was deemed exempt from review by an institutional review board.

Due to the wide variety of authors, contributors, and editors involved in the creation of these materials, the unclear edit history of the documents, and the presence of former students who no longer maintain active institutional email addresses, informed consent from all contributors could not feasibly be obtained. Ethical considerations were therefore evaluated using the framework for internet-mediated research proposed by Eysenbach and Till [22].

Within this framework, the study was determined to be ethically appropriate. Regarding intrusiveness, the project consisted solely of passive analysis of existing materials and did not involve interaction with members of the online community. In terms of perceived privacy, the study guides are hosted on a shared institutional drive that has been openly used by entire classes of Uniformed Services University of the Health Sciences medical students for approximately 5 years, with each class comprising roughly 100170 students.

With respect to vulnerability, only minimal considerations apply, as participants represent adult trainees within an academic military medical program. Potential harm was assessed as negligible because the analyzed materials are purely academic study guides and do not contain sensitive personal information. Accordingly, informed consent was waived.

To protect confidentiality, the authors attest that all downloaded materials were reviewed prior to analysis to ensure that no personally identifiable information was present. Finally, regarding intellectual property considerations, the study guides were reviewed, and no copyrighted material was identified within the documents.

Results

Overall Performance

Detailed trial-level results are listed below in Table 1. Each trial consists of 121 responses, and omitted responses represent instances in which a model did not produce a definitive answer during a trial.

While ChatGPT-4o Mini achieved a peak single-trial accuracy of 85.9% (104/121) in trials 2 and 3, the overall question-level (majority-consensus) accuracy was 84.30% (102/121). NLM w/ Notes had an overall accuracy rate of 74.38% (90/121); NLM w/o Notes had an overall accuracy rate of 66.94% (81/121). Finally, Gemini 1.5 Flash had an overall accuracy rate of 64.46% (78/121; Figure 2).

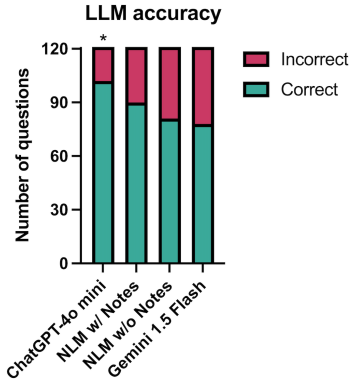
Table 1. By-trial breakdown of correct, nonomitted incorrect, and omitted data for each LLM tested in an observational, comparative performance evaluation of LLMs answering United States Medical Licensing Examination (USMLE) step 1 dermatology questions conducted during December 2024.

LLM[a] and trial number	Correct, n (%)	Nonomitted incorrect, n (%)	Omitted, n (%)
ChatGPT-4o Mini			
1	102 (84.3)	18 (14.9)	1 (0.826)
2	104 (85.9)	17 (14.0)	0 (0)
3	104 (85.9)	17 (14.0)	0 (0)
NLM[b] w/ Notes			
1	91 (75.2)	28 (23.1)	2 (1.65)
2	92 (76.0)	29 (24.0)	0 (0)
3	91 (76.0)	30 (24.8)	0 (0)
NLM w/o Notes			
1	83 (68.6)	36 (29.8)	2 (1.65)
2	80 (66.1)	22 (18.2)	19 (15.7)
3	82 (67.8)	22 (18.2)	17 (14.0)
Gemini 1.5 Flash			
1	79 (65.3)	39 (32.2)	3 (2.48)
2	79 (65.3)	39 (32.2)	3 (2.48)
3	79 (65.3)	41 (33.9)	1 (0.826)

[a]LLM: large language model.
[b]NLM: NotebookLM.

ChatGPT-4o Mini scored significantly higher than NLM w/ Notes (McNemar χ^2_1=7.2; BH-adjusted P=.018), NLM w/o Notes (McNemar χ^2_1=14.2; BH-adjusted P=.002), and Gemini 1.5 Flash (McNemar χ^2_1=16.9; BH-adjusted P<.001).

Figure 2. Overall performance of each tested large language model (LLM), assessing the majority answer given out of three trials and comparing it to the answer key in an observational, comparative performance evaluation of LLMs answering United States Medical Licensing Examination (USMLE) step 1 dermatology questions conducted during December 2024. ChatGPT-4o Mini displayed a significantly higher accuracy rate (*Benjamini-Hochberg adjusted P<.05) than all other LLMs tested. NLM: NotebookLM.



Performance of the LLM by Question Difficulty

ChatGPT-4o Mini had a 100% (7/7) accuracy rate for "1 Hammer" questions, a 92.11% (35/38) accuracy rate for "2 Hammer" questions, an 85.42% (41/48) accuracy rate for "3 Hammer" questions, a 73.08% (19/26) accuracy rate for "4 Hammer" questions, and a 0% (0/2) accuracy rate for "5 Hammer" questions.

NLM w/ Notes had a 100% (7/7) accuracy rate for "1 Hammer" questions, an 89.47% (34/38) accuracy rate for "2 Hammer" questions, a 70.83% (34/48) accuracy rate for "3 Hammer" questions, a 57.69% (15/26) accuracy rate for "4 Hammer" questions, and a 0% (0/2) accuracy rate for "5 Hammer" questions.

NLM w/o Notes had a 100% (7/7) accuracy rate for "1 Hammer" questions, a 71.05% (27/38) accuracy rate for "2 Hammer" questions, a 62.5% (30/48) accuracy rate for "3 Hammer" questions, a 69.23% (18/26) accuracy rate for "4 Hammer" questions, and a 0% (0/2) accuracy rate for "5 Hammer" questions.

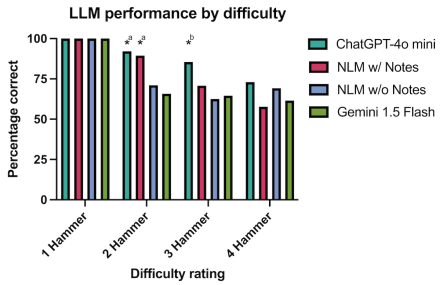
Gemini 1.5 Flash had a 100% (7/7) accuracy rate for "1 Hammer" questions, a 65.79% (25/38) accuracy rate for "2 Hammer" questions, a 64.58% (31/48) accuracy rate for "3

Hammer" questions, a 61.54% (16/26) accuracy rate for "4 Hammer" questions, and a 0% (0/2) accuracy rate for "5 Hammer" questions.

ChatGPT-4o Mini and NLM w/ Notes had a significantly higher overall accuracy than NLM w/o Notes (ChatGPT-4o Mini vs NLM w/o Notes: McNemar χ^2_1=6.4, BH-adjusted P=.02; NLM w/ Notes vs NLM w/o Notes: McNemar χ^2_1=7.0, BH-adjusted P=.02) and Gemini 1.5 Flash (ChatGPT-4o Mini vs Gemini 1.5 Flash: McNemar χ^2_1=10.0, BH-adjusted P=.009; NLM w/ Notes vs Gemini 1.5 Flash: McNemar χ^2_1=9.0, BH-adjusted P=.009) in the "2 Hammer" category.

For the "3 Hammer" difficulty category, ChatGPT-4o Mini scored significantly higher than NLM w/o Notes (McNemar χ^2_1=9.3; BH-adjusted P=.009) and Gemini 1.5 Flash (McNemar χ^2_1=8.3; BH-adjusted P=.01). All the LLM aggregate scores were 100% for "1 Hammer" questions and 0% for "5 Hammer" questions (Figure 3).

Figure 3. Comparison of large language model (LLM) overall accuracy rates by difficulty level, with "1 Hammer" being the least difficult and "5 Hammer" being the most difficult, in an observational, comparative performance evaluation of LLMs answering United States Medical Licensing Examination (USMLE) step 1 dermatology questions conducted during December 2024. Results for "5 Hammer" questions are not displayed due to all models scoring 0%. [a] Within the category of "2 Hammer" questions: ChatGPT-4o Mini and NLM w/ Notes did not differ significantly from each other (Benjamini–Hochberg adjusted P<.05). However, both of these LLMs performed significantly better than NLM w/o Notes and Gemini 1.5 Flash. [b] Within the category of "3 Hammer" questions: ChatGPT-4o Mini performed significantly better than NLM w/o Notes and Gemini 1.5 Flash; there was no significant difference from NLM w/ Notes. NLM: NotebookLM.

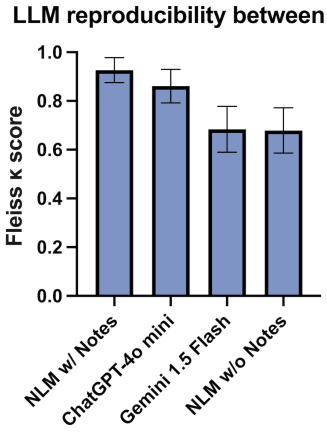


With regard to the difficulty distribution of omissions of NLM w/o Notes, a chi-square analysis found no significant difference in the omission rate between the question difficulty levels: χ^2_3=1.9; P=.76.

Reproducibility

We used a Fleiss ϰ to measure the reproducibility of answer selection (A, B, C, D, E, F, G, and X) chosen by each LLM over their respective 3 trials. NLM w/ Notes had a Fleiss ϰ of 0.927 (95% CI 0.875-0.978), ChatGPT-4o Mini had a ϰ statistic of 0.861 (95% CI 0.792-0.929), Gemini 1.5 Flash had a ϰ statistic of 0.684 (95% CI 0.590-0.778), and NLM w/o Notes had a ϰ statistic of 0.679 (95% CI 0.586-0.772). NLM w/ Notes and ChatGPT-4o Mini had significantly higher reproducibility than NLM w/o Notes and Gemini 1.5 Flash (Figure 4).

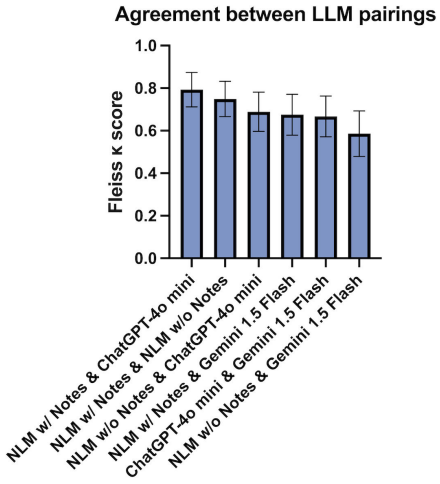
Figure 4. Calculated Fleiss ϰ scores demonstrating the capability of large language models (LLMs) to give concordant answer choice responses between repeat trials in an observational, comparative performance evaluation of LLMs answering United States Medical Licensing Examination (USMLE) step 1 dermatology questions conducted during December 2024. 95% CIs are displayed alongside Fleiss ϰ scores. NLM: NotebookLM.



Agreement Between the LLMs

Finally, we assessed the presence of cross-AI "agreement" using a Fleiss ϰ, comparing the answer selection (A, B, C, D, E, F, G, and X) of all the trials of two LLMs (comparing 6 trials in total) to determine the pairwise agreement between different LLMs across the answered questions (excluding questions where at least 1 trial had an omission). Pairwise Fleiss ϰ scores were 0.793 (95% CI 0.7119-0.873) between NLM w/ Notes and ChatGPT-4o Mini, 0.749 (95% CI 0.662-0.835) for NLM w/ Notes and NLM w/o Notes, 0.689 (95% CI 0.596-0.782) for NLM w/o Notes and ChatGPT-4o Mini, 0.675 (95% CI 0.580-0.770) for NLM w/ Notes and Gemini 1.5 Flash, 0.667 (95% CI 0.572-0.763) for ChatGPT-4o Mini and Gemini 1.5 Flash, and 0.586 (95% CI 0.486-0.686) for NLM w/o Notes and Gemini 1.5 Flash. The only statistically significant difference in agreement was between the NLM w/ Notes and ChatGPT-4o Mini pairing and the NLM w/o Notes and Gemini 1.5 Flash pairing (Figure 5).

Figure 5. Calculated Fleiss κ scores demonstrating the pairwise capability of different large language models (LLMs) to give concordant answer choice responses within all 6 of their trials in an observational, comparative performance evaluation of LLMs answering United States Medical Licensing Examination (USMLE) step 1 dermatology questions conducted during December 2024. 95% CI are displayed alongside Fleiss κ scores. NLM: NotebookLM.



Discussion

ChatGPT-4o Mini had the highest overall rate of accuracy, demonstrated significantly higher accuracy on "2 Hammer" and "3 Hammer" questions, and exhibited significantly higher intertrial reproducibility levels than NLM w/o Notes and Gemini 1.5 Flash. NLM w/ Notes exhibited fewer omissions, higher rates of intertrial reproducibility, and higher accuracy on "2 Hammer" questions compared to NLM w/o Notes. NLM w/ Notes had a higher accuracy than NLM w/o Notes and Gemini 1.5 Flash; however, the difference did not meet the threshold for significance. Among the models, the highest agreement rates were observed between NLM w/ Notes and ChatGPT-4o Mini, which were significantly higher than those of NLM w/o Notes and Gemini 1.5 Flash. There were no significant differences between NLM w/o Notes and Google Gemini 1.5 Flash in terms of accuracy or reproducibility.

The omission rate was disproportionately high for NLM w/o Notes; 22 of 121 questions had at least one trial-level omission across the 3 trials, compared to fewer than 7 of 121 questions for other models (Multimedia Appendix 1). There was no significant trend in the difficulty rating for omitted questions of NLM w/o Notes; however, sensitivity analysis (Multimedia Appendix 1) excluding omissions showed improved performance in NLM w/o Notes (from 81/121, 66.9% to 77/99, 77.8% overall accuracy), reaching the threshold where NLM w/o Notes had statistically significant differences from Gemini 1.5 Flash in overall accuracy. This pattern suggests that omissions occurred preferentially in questions that NLM's foundational LLM was less able to solve.

The overall performance of LLMs at the time of this study was similar to that reported in other published papers. A March 2024 paper on ChatGPT 4 showed an overall accuracy rate of 85.7% (1157/1350) on general step 1 and step 2 AMBOSS questions, similar to our overall accuracy rate of 84.30% (102/121) for ChatGPT-4o Mini [23]. Our results provide further evidence for the increased performance of ChatGPT-4o Mini compared to its previous version's accuracy rate on general AMBOSS step 1 questions of 44% (44/100) in December 2022 [24] and demonstrate the relative capabilities of Gemini 1.5 Flash and NLM. Data on Gemini 1.5 Flash's prior performance on AMBOSS questions were not found.

The RAG technique used in NLM involves splitting documents into indexed sections and retrieving sections for use in responses based on their similarity to the query. It has been shown to reduce hallucinations in LLMs due to the grounding of answers within source material [25]. This may account for the significantly higher intertrial reproducibility seen in NLM w/ Notes compared to NLM w/o Notes observed in both primary and sensitivity analyses, since the RAG algorithms within NLM may choose the same source sections between trials (regardless of their appropriateness) to use as the basis for an answer.

A similar mechanism involving RAG excerpt selection may also explain the trends in accuracy across question difficulty levels observed for NLM w/ Notes.

The increase in accuracy on "2 Hammer" questions for NLM when it was provided with notes could possibly be explained by "2 Hammer" questions having text aligned with classic keywords of disease scripts. Within the culture of medicine, providers are regularly trained to use certain keywords that concisely convey diagnostic concepts. For example, referring to subglottic narrowing as the "steeple sign" to point toward the diagnosis of croup [26]. The relatively low difficulty of "2 Hammer" questions may have produced vignettes whose disease descriptions closely aligned with keywords present in the student-created sources. Furthermore, follow-up questions to these vignettes might have included more "lower order" questions that rely on pure factual recall [27]. For example, 1 "2 Hammer" question stated that a patient had skin lesions with "a dusky center with a lighter ring around them," which was a close allusion to the target sign found in erythema multiforme, and asked, "Which of the following is the most likely diagnosis?" which aligned with a pure factual recall of erythema multiforme.

In contrast, higher-difficulty questions may present with more complex constellations of symptoms, descriptions not as well characterized by keywords, or "higher-order" questions that require a conceptual understanding of more distant concepts. For example, one "4 Hammer" question stated that a patient with mild fever had "multiple vesicles, flaccid transparent bullae that contain clear yellow fluid, and brown crusts on her chest and upper extremities." Application of a shear force to the surrounding unaffected skin does not cause sloughing, with the main question asking about the underlying cause of the condition. Here, the patient's presentation did not correspond to a single concise diagnostic keyword; any entity responding to this question had to use context clues to know the etiology (bullous impetigo) and apply that

to knowledge of bacterial virulence factors (exfoliative toxin A). It is possible that for the higher difficulty questions, the complex question nature may have led NLM RAG algorithms to choose the wrong source material sections, as evidenced by the decrease in accuracy for NLM w/ Notes for the "4 Hammer" questions. Although the decrease was not statistically significant, the observed phenomenon and underlying RAG quality are still areas that could be explored.

This paper had several limitations and areas for future research. One limitation was related to the variable quality of the uploaded sources. This was part of the experimental design, as student-derived notes were provided to NLM with the understanding that they were not expertly curated and might contain extraneous, incomplete, or even incorrect information—similar to any student notes that could be uploaded to a source-based LLM focused on CLT-informed teaching on an individual level. It appears that when working with student-generated sources, an educationally focused LLM may need to progress in striking a balance between using the sources and using baseline encoded reasoning to shape answers. Furthermore, such an LLM would be strengthened by the ability to check uploaded material and address misconceptions or gaps in knowledge when they appear.

Another limitation of this study was the design of having three repeated queries before a question was considered "omitted." Here, our aim was to maximize AI responsiveness, and the formative nature of this study prioritized evaluating a model's ability to ultimately generate an answer rather than penalizing initial nondefinitive responses. This approach may have underestimated the true frequency of omissions and could have subtly influenced internal model states despite attempts to refresh interactions between questions. Combined, the 3 repeated queries may have artificially inflated accuracy rates, especially in models with higher omission frequencies. Future studies may benefit from comparing single-query and multiquery omission thresholds to better characterize LLM responsiveness.

The narrow selection of the question set limited the generalizability of the results in other domains of medicine. Furthermore, image-dependent knowledge is crucial in dermatology, and the exclusion of image-based questions limits the application to real-world dermatologic reasoning and comprehensive dermatologic competence. NLM did not support image uploads to its queries at the time of this study.

Another limitation was the use of a blank piece of paper as a negative control, since the paper may have served as a source-grounding constraint condition where the NLM would omit answers that did not match its source. As mentioned before, sensitivity analysis showed some evidence that answer omissions occurred preferentially in questions that the AI could not solve, suggesting that the model's architecture prioritized source fidelity over response rate. The use of a blank piece of paper as a negative control may have decreased the accuracy rate of the NLM w/o Notes, since its baseline AI could possibly have correctly answered some questions that were omitted. Future studies could introduce expert-curated sources as a "positive control"; however, additional difficulties lie in choosing a representative dermatology-specific authoritative source and uploading extensive copyrighted material.

Further limitations included the attempts to "reset" an LLM through refreshes and new instances. Although the LLMs tested deny training answers based on specific user chats, there still may have been a subtle influence of unclear magnitude or direction on later answers. Such a phenomenon was not avoidable in our trials.

Overall, this study provides data on the accuracy of contemporary LLMs, yielding accuracy rates similar to those found in the existing literature. Our findings on reproducibility and accuracy by question difficulty suggest that NLM's RAG algorithm significantly stabilizes model output, likely by consistently retrieving the same source text excerpts across repeated trials.

Furthermore, the lack of accuracy improvement and the decrease in performance of NLM w/ Notes when exposed to more difficult questions suggest that RAG's statistically favored excerpts from source texts may not match the excerpts needed to answer a question properly—this is in line with previously characterized phenomena of "retrieval failure" found in LLMs that use RAG [28]. Further evidence for "retrieval failure" was seen in the significantly increased reproducibility in the NLM w/ Notes condition, suggesting the involvement of similar source excerpts across trials guiding similar answers.

Previous studies have assessed the quality of source-based LLMs when given expert-curated material [17]. This has many uses in areas like medical diagnosis, clinical decision-making, or areas that require precise, expert-level performance. However, few have considered how an LLM performs when intentionally given suboptimal material—material that better captures what might be uploaded to one of these tools by an average student. The observed retrieval failures, likely exacerbated by the noncurated nature of the notes, provide insight into characteristics that would make a stronger student-level CLT-grounded educational LLM—one that can strike a sophisticated balance between using an imperfect source and drawing from its own encoded knowledge to meaningfully challenge a student when there are misconceptions, gaps, or even biases embedded within such sources. By accounting for imperfections within provided sources, an educational LLM may bridge gaps in knowledge and serve as an efficient guide for students along the path toward clinical mastery.

Acknowledgments

The authors wish to thank the Uniformed Services University contributors for providing the inputted exam study guides used in this investigation. Generative artificial intelligence was not used in the ideation or writing process for this manuscript.

Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

Disclaimer

The opinions and assertions expressed herein are those of the authors and do not reflect the official policy or position of the Uniformed Services University of the Health Sciences or the Department of Defense.

Data Availability

The datasets generated or analyzed during this study, including the AMBOSS link to the question set, R code, and data sheets, are available from the corresponding author upon reasonable request.

Authors' Contributions

FJML wrote, edited, and formally analyzed the data in this manuscript. SC supervised, assisted in the conceptualization of the original study design, and edited the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Sensitivity analysis.
[PDF File (Adobe File), 684 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Statistical code and output for primary analyses.
[PDF File (Adobe File), 88 KB-Multimedia Appendix 2]

Checklist 1

STROBE checklist.
[PDF File (Adobe File), 127 KB-Checklist 1]

References

1. Tran C, Hryciw BN, Moore SW, Chaput A, Seely AJE. Perceptions and use of generative artificial intelligence in medical students: a multicenter survey. J Med Educ Curric Dev. 2025;12:23821205251391969. [doi: 10.1177/23821205251391969] [Medline: 41181167]
2. Ramos B, Condotta R. Enhancing learning and collaboration in a unit operations course: using AI as a catalyst to create engaging problem-based learning scenarios. J Chem Educ. Aug 13, 2024;101(8):3246-3254. [doi: 10.1021/acs.jchemed.4c00244]
3. Kejingyun S, Mingjun R. Randomized controlled study on the impact of problem-based learning combined with large language models on critical thinking skills in nursing students. Nurse Educ. 2025;50(4):216-220. [doi: 10.1097/NNE.0000000000001879] [Medline: 40261697]
4. Huang C, Zhong Y, Li Y, et al. Enhancing student reading performance through a personalized two-tier problem-based learning approach with generative artificial intelligence. Humanit Soc Sci Commun. 2025;12(645). [doi: 10.1057/s41599-025-04919-4]
5. Gkintoni E, Antonopoulou H, Sortwell A, Halkiopoulos C. Challenging cognitive load theory: the role of educational neuroscience and artificial intelligence in redefining learning efficacy. Brain Sci. Feb 15, 2025;15(2):203. [doi: 10.3390/brainsci15020203] [Medline: 40002535]
6. Albrecht-Crane C. Thinking smarter, not harder? Google NotebookLM's misalignment problem in education. Presented at: SIGDOC '25: Proceedings of the 43rd ACM International Conference on Design of Communication; Oct 24-25, 2025:121-127; Lubbock, TX. [doi: 10.1145/3711670.3764628]
7. NotebookLM. URL: https://notebooklm.google/ [Accessed 2026-06-03]
8. Sweller J. Cognitive load theory and individual differences. Learn Individ Differ. Feb 2024;110(1):102423. [doi: 10.1016/j.lindif.2024.102423]
9. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. J Med Internet Res. Apr 5, 2024;26:e52935. [doi: 10.2196/52935] [Medline: 38578685]
10. Wu K, Wu E, Wei K, et al. An automated framework for assessing how well LLMs cite relevant medical references. Nat Commun. Apr 16, 2025;16(1):3615. [doi: 10.1038/s41467-025-58551-6] [Medline: 40240349]

11. Hardinges J, Simperl E, Shadbolt N. We must fix the lack of transparency around the data used to train foundation models. Harv Data Sci Rev. 2023. [doi: 10.1162/99608f92.a50ec6e6]
12. Mangion SE, Phan TA, Zagarella S, Cook D, Ganda K, Maibach HI. Medical school dermatology education: a scoping review. Clin Exp Dermatol. Jun 5, 2023;48(6):648-659. [doi: 10.1093/ced/llad052] [Medline: 36753386]
13. Janodia R, Nguyen H, Fitzhugh VA, Traba C, Chen S, Grachan JJ. Addressing visual learning equity in undergraduate dermatology education: skin color representation across dermatology lecture images at Rutgers New Jersey Medical School. J Natl Med Assoc. Feb 2025;117(1):74-79. [doi: 10.1016/j.jnma.2025.01.010] [Medline: 39952847]
14. Adelekun A, Onyekaba G, Lipoff JB. Skin color in dermatology textbooks: an updated evaluation and analysis. J Am Acad Dermatol. Jan 2021;84(1):194-196. [doi: 10.1016/j.jaad.2020.04.084] [Medline: 32335181]
15. Omar M, Sorin V, Agbareia R, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. Int J Equity Health. Feb 26, 2025;24(1):57. [doi: 10.1186/s12939-025-02419-0] [Medline: 40011901]
16. Rao AS, Kim J, Mu A, et al. Synthetic medical education in dermatology leveraging generative artificial intelligence. NPJ Digit Med. May 4, 2025;8(1):247. [doi: 10.1038/s41746-025-01650-x] [Medline: 40320492]
17. Perkins G, Anderson NW, Spies NC. Retrieval-augmented generation salvages poor performance from large language models in answering microbiology-specific multiple-choice questions. J Clin Microbiol. Mar 12, 2025;63(3):e0162424. [doi: 10.1128/jcm.01624-24] [Medline: 39932275]
18. Question difficulty. AMBOSS Support. URL: https://support.amboss.com/hc/en-us/articles/360035679652-Question-difficulty [Accessed 2026-02-02]
19. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. Oct 16, 2007;4(10):e296. [doi: 10.1371/journal.pmed.0040296] [Medline: 17941714]
20. Signorell A. DescTools: tools for descriptive statistics. CRAN (Comprehensive R Archive Network). Mar 28, 2025. URL: https://cran.r-project.org/package=DescTools [Accessed 2026-06-03]
21. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. Apr 1960;20(1):37-46. [doi: 10.1177/001316446002000104]
22. Eysenbach G, Till JE. Ethical issues in qualitative research on internet communities. BMJ. Nov 10, 2001;323(7321):1103-1105. [doi: 10.1136/bmj.323.7321.1103] [Medline: 11701577]
23. Funk PF, Hoch CC, Knoedler S, et al. ChatGPT's response consistency: a study on repeated queries of medical examination questions. Eur J Investig Health Psychol Educ. Mar 8, 2024;14(3):657-668. [doi: 10.3390/ejihpe14030043] [Medline: 38534904]
24. Gilson A, Safranek CW, Huang T, et al. 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 Med Educ. Feb 8, 2023;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
25. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. PLOS Digit Health. Jun 2025;4(6):e0000877. [doi: 10.1371/journal.pdig.0000877] [Medline: 40498738]
26. Chan MW, Eppich WJ. The keyword effect: a grounded theory study exploring the role of keywords in clinical communication. AEM Educ Train. 2019;4(4):403-410. [doi: 10.1002/aet2.10424] [Medline: 33150283]
27. Tofade T, Elsner J, Haines ST. Best practice strategies for effective use of questions as a teaching tool. Am J Pharm Educ. Sep 12, 2013;77(7):155. [doi: 10.5688/ajpe777155] [Medline: 24052658]
28. Barnett S, Kurniawan S, Thudumu S, Brannelly Z, Abdelrazek M. Seven failure points when engineering a retrieval augmented generation system. Presented at: CAIN '24: Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI; Apr 14-15, 2024:194-199; Lisbon, Portugal. [doi: 10.1145/3644815.3644945]

Abbreviations:

AI : artificial intelligence
BH: Benjamini-Hochberg
CLT: cognitive load theory
LLM: large language model
NLM: NotebookLM
RAG: retrieval-augmented generation
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology
USMLE: United States Medical Licensing Examination

Edited by Javad Sarvestan; peer-reviewed by Polina Shilo, Sadhasivam Mohanadas, Toshimune Ito; submitted 18.Nov.2025; final revised version received 11.Mar.2026; accepted 12.Mar.2026; published 25.Jun.2026

Please cite as:
Lin FJM, Cho S
Evaluating Source-Based Large Language Models for Preclinical Dermatology Education: Comparative Study
JMIR Form Res 2026;10:e88008
URL: https://formative.jmir.org/2026/1/e88008
doi: 10.2196/88008