Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative Benchmarking of ChatGPT, Claude, and Grok on Persian General Surgery Promotion Examinations With and Without an Uploaded Surgical Reference Package

Shervan Family[1]
Islamic Azad University, Tehran Medical Sciences Branch
Sina Raissi Dehkordi[1]
Islamic Azad University, Tehran Medical Sciences Branch
Parmida Pahlavanzadeh[2]
Parmiida.p.z@gmail.com

Islamic Azad University, Tehran Medical Sciences Branch

Article

Keywords:

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11098215/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Large language models are increasingly used in medical education, yet their observed agreement with answer keys on locally developed, non-English surgical examinations remains incompletely characterized.

Methods

This comparative benchmark evaluated ChatGPT (Luna High), Claude Opus 5, and Grok 4/5 on 456 original Persian four-option questions from the 2024 (1403 Persian calendar), 2025 (1404 Persian calendar), and 2026 (1405 Persian calendar) Iranian General Surgery Residency Promotion Examinations. Each annual question file was submitted once to each model in two separate user-facing workflows: question file only and question file plus an uploaded surgical reference package. Responses were scored against the official Sanjesh answer keys using a strict single-option rule.

Results

With the uploaded reference package, aggregate answer-key agreement was 71.27% for ChatGPT, 75.44% for Grok, and 80.04% for Claude. In the question-file-only workflow, the corresponding values were 67.11%, 72.59%, and 79.82%. Claude had the highest observed aggregate agreement in both workflows.

Conclusion

In this single-run, web-enabled benchmark, Claude showed the highest observed aggregate agreement with the official answer keys. The uploaded reference package was associated with a positive change in ChatGPT agreement across all three years, whereas the direction of change varied for Grok and Claude. These findings are descriptive and do not establish a general model ranking or a causal effect of document upload.

Introduction

Evidence about LLM performance on Persian, file-based general surgery promotion examinations is limited. These examinations use locally developed Persian-language questions and an official answer key; evidence from English-language medical benchmarks cannot be assumed to transfer directly to this workflow [1–4].

Observed agreement with medical examination keys varies across model configurations, specialties, question formats, and languages [5,6,9–11,13,14]. In this study, agreement with the official key is an examination outcome; it is not evidence of independent clinical reasoning or readiness for unsupervised clinical practice.

Prior surgical LLM evaluations include Korean and US general surgery examinations [7,8]. A recent Iranian study evaluated item-level Persian surgical subspecialty board questions [15]. Neither approach examined the user-facing workflow evaluated here, in which an annual Persian promotion-examination file was uploaded as a whole, either alone or with an uploaded surgical document package.

We evaluated the observed agreement of three commercial LLM configurations with the official answer keys when annual Persian examination files were uploaded either alone or alongside a predefined surgical document package. The primary aim was descriptive and was limited to answer-key agreement within this user-facing workflow.

This benchmark does not assess clinical competence, diagnostic safety, or autonomous clinical decision-making.

Methods

Study design and question corpus

This comparative benchmark study assessed the observed agreement of ChatGPT (Luna High), Claude Opus 5, and Grok 4/5 with the official answer keys for the Iranian General Surgery Residency Promotion Examination, a written assessment for general surgery residents. Examination sets from 2024 (1403 Persian calendar), 2025 (1404 Persian calendar), and 2026 (1405 Persian calendar) of the Persian calendar were included. All questions were presented in their original Persian wording and were four-option multiple-choice questions with one official keyed answer. No included item required interpretation of an image, figure, graph, or other non-textual content.

Each annual set contained 152 questions, including two ethics items at the end of the examination. No text-based questions were excluded, yielding 456 questions for the descriptive analysis. For each year, the original examination questions were submitted together, in their original order, as a single file rather than as separate item-level prompts. The benchmark therefore assessed document-level examination answering rather than isolated single-question performance.This study did not involve direct intervention with human participants

Model access and evaluation conditions

The three model configurations were accessed through their respective official user-facing applications during July–August 2026 (Mordad 1405). All interactions were initiated in separate new conversations,

and internet access was enabled throughout the evaluation; therefore, the question-file-only workflow was not a closed-book condition. Each annual examination file was submitted once to each model in each workflow. The two workflows were conducted in separate new sessions to reduce carryover from previously uploaded materials.

The uploaded reference package consisted of Schwartz's Principles of Surgery, 11th edition (2019); Sabiston Textbook of Surgery, 21st edition (2022); Maingot's Abdominal Operations, 13th edition (2019); Zollinger's Atlas of Surgical Operations, 11th edition (2021) [16–19]; and selected articles from Surgical Clinics of North America.

Prompt

The same Persian prompt was used across all models and both workflows. The original Persian wording and one English translation are provided in Appendix 1; the translation was included for reporting transparency only and was not supplied to the models. The presence of the uploaded surgical document package was the intended workflow difference. Because document consultation was not directly observable and the prompt did not instruct models to cite or use the package, the study cannot establish an effect of reference use.

Outcome assessment

Two authors independently compared each recorded model response with the official answer key for the corresponding examination. The official answer keys were obtained from the Iranian National Organization for Educational Testing (Sanjesh). A response was classified as correct only when it identified one unambiguous answer that matched the official key. Responses containing multiple incompatible answers, no definitive answer, or an answer that did not match the official key were classified as incorrect under the strict examination-style scoring rule. Disagreements were resolved by re-examining the recorded model output against the official answer key.

During quality control, the reference-package Grok run for the 2024 (1403 Persian calendar) examination was repeated after an unexpected initial result. The repeated output was independently scored by the same two reviewers and is reported as a replication observation for that model-year-workflow combination, not as a fixed model characteristic.

Statistical analysis

Analyses were descriptive. For each model-year-workflow combination, answer-key agreement was calculated as the number of correct responses divided by 152 and was summarized as counts and percentages by examination year and workflow. Paired item-level outcomes were not retained; therefore, paired between-model and between-condition comparisons were not performed. The analysis was limited to the available aggregate results from single runs.

Reporting was guided by LLM-specific transparency resources, including MI-CLEAR-LLM, TRIPOD-LLM, and the DEAL checklist [20–22].

Results

Descriptive accuracy

Table 1 and Fig. 1 summarize key-concordant responses by model, examination year, and workflow.

Table 1
Agreement with the official answer key by model, examination year, and document-upload workflow

Examination year	Model	With uploaded reference package, n/N (%)	Question file only, n/N (%)
2026 (1405 Persian calendar)	ChatGPT	111/152 (73.03)	110/152 (72.37)
2026 (1405 Persian calendar)	Grok	120/152 (78.95)	125/152 (82.24)
2026 (1405 Persian calendar)	Claude	130/152 (85.53)	129/152 (84.87)
2025 (1404 Persian calendar)	ChatGPT	119/152 (78.29)	104/152 (68.42)
2025 (1404 Persian calendar)	Grok	144/152 (94.74)	109/152 (71.71)
2025 (1404 Persian calendar)	Claude	110/152 (72.37)	130/152 (85.53)
2024 (1403 Persian calendar)	ChatGPT	95/152 (62.50)	92/152 (60.53)
2024 (1403 Persian calendar)	Grok	80/152 (52.63)	97/152 (63.82)
2024 (1403 Persian calendar)	Claude	125/152 (82.24)	105/152 (69.08)

Values are reported as key-concordant responses/152 questions (%). Examination years are shown as Gregorian year (Persian calendar year).

Table 2
Aggregate agreement with the official answer key across the three examination years

Model	With uploaded reference package (n/N; %)	Question file only (n/N; %)
ChatGPT	325/456 (71.27%)	306/456 (67.11%)
Grok	344/456 (75.44%)	331/456 (72.59%)
Claude	365/456 (80.04%)	364/456 (79.82%)

Reference-supported condition

With the uploaded surgical document package, aggregate answer-key agreement across the three years was 325/456 (71.27%) for ChatGPT, 344/456 (75.44%) for Grok, and 365/456 (80.04%) for Claude. Claude had the highest observed aggregate agreement in this workflow. The highest single-year agreement was observed for Grok in the 2025 (1404 Persian calendar) examination, with 144/152 key-concordant responses (94.74%).

The reference-package Grok evaluation for the 2024 (1403 Persian calendar) examination was repeated as a quality-control check. The repeated output agreed with the official answer key for 80/152 questions (52.63%).

Question-file-only condition

Without an uploaded surgical document package, aggregate answer-key agreement across the three years was 306/456 (67.11%) for ChatGPT, 331/456 (72.59%) for Grok, and 364/456 (79.82%) for Claude. Claude again had the highest observed aggregate agreement. The highest single-year agreement in this workflow was observed for Claude in the 2025 (1404 Persian calendar) examination, with 130/152 key-concordant responses (85.53%).

Comparison between conditions

Observed changes after upload of the reference package were not uniform across years or models (Table 1). ChatGPT showed higher agreement with the package in 2025 (1404 Persian calendar) and 2024 (1403 Persian calendar); Grok showed higher agreement in 2025 (1404 Persian calendar) but lower agreement in 2026 (1405 Persian calendar) and 2024 (1403 Persian calendar); and Claude showed higher agreement in 2026 (1405 Persian calendar) and 2024 (1403 Persian calendar) but lower agreement in 2025 (1404 Persian calendar). These are descriptive patterns only.

Discussion

Claude had the highest observed aggregate answer-key agreement in both workflows (Table 2). This is a descriptive finding from one web-enabled run per model-year-workflow combination and should not be interpreted as evidence of a stable general model ranking.

Aggregate values concealed year-specific variation. Grok produced the highest single-year value in the reference-package 2025 (1404 Persian calendar) workflow, whereas Claude had the highest observed agreement in both workflows for 2026 (1405 Persian calendar) and in the question-file-only workflow for 2025 (1404 Persian calendar) (Table 1).

Availability of the uploaded document package coincided with higher observed agreement for ChatGPT across the three examination years, whereas the direction of change varied for Grok and Claude. Because document use was not directly observable, this pattern should not be interpreted as a causal effect of reference consultation.

The available data cannot distinguish among potential explanations for these patterns, including differences in attachment processing, discordance between an uploaded source and the official key, and variation in question mix across examination years.

The low Grok result in the reference-package 2024 (1403 Persian calendar) workflow should be interpreted cautiously. The repeated run agreed with the official key for 80/152 questions (52.63%); this replication observation does not by itself establish within-model variability or a fixed model characteristic.

Few published evaluations have examined Persian, document-level general surgery promotion examinations. The present study contributes a benchmark based on original Persian examination files, official answer keys, and two user-facing document-upload workflows [7,8,15].

Strengths include the three consecutive examination years, retention of all text-based questions, use of the same Persian prompt across workflows, and scoring against official answer keys. Each annual file was submitted in its original order as one document rather than as 152 separate prompts.

Interpretation is limited by single runs, aggregate rather than retained item-level outcomes, internet-enabled consumer applications, incomplete reproducibility information for platform behavior, and the absence of blinded scorer agreement statistics or expert adjudication of potential key-versus-clinical disagreements. The prompt did not direct models to use the uploaded package, and the initial Grok output and complete file manifest should be retained and reported in supplementary material. Permissions for examination materials and uploaded references also require confirmation before submission.

Future work should retain item-level response logs, record exact model configurations and run dates, prespecify the handling of technical failures and repeats, report inter-rater agreement, and classify

questions by topic and cognitive task. Repeated runs and paired item-level data would permit more informative comparisons.

Conclusion

In this single-run, web-enabled benchmark, Claude showed the highest observed aggregate agreement with the official answer keys in both document-upload conditions. The availability of an uploaded surgical document package was associated with a positive change in ChatGPT's observed agreement across the three examination years, whereas the direction of change varied for Grok and Claude. Because item-level paired outcomes and repeated runs were unavailable, these findings are descriptive and do not establish a general model ranking or a causal effect of the uploaded reference package.

Declarations

Author Contribution

S.F., S.R.D., and P.P. contributed equally to the study's conception, design, data collection, analysis, and manuscript preparation. All authors read and approved the final manuscript.

Data Availability

The data supporting the findings of this study are available from the corresponding author on reasonable request.

References

1. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25** (1), 44–56. 10.1038/s41591-018-0300-7 (2019).
2. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29** (8), 1930–1940. 10.1038/s41591-023-02448-8 (2023).
3. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, 141. 10.1038/s43856-023-00370-1 (2023).
4. Rajpurkar, P. et al. AI in health and medicine. *Nat. Med.* **28** (1), 31–38. 10.1038/s41591-021-01614-0 (2022).
5. Kung, T. H. et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit. Health*. **2** (2), e0000198. 10.1371/journal.pdig.0000198 (2023).
6. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620** (7972), 172–180. 10.1038/s41586-023-06291-2 (2023).

7. Oh, N., Choi, G-S. & Lee, W. Y. ChatGPT goes to the operating room: Evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Ann. Surg. Treat. Res.* **104** (5), 269–273. 10.4174/astr.2023.104.5.269 (2023).
8. Tran, C. G., Chang, J., Sherman, S. K. & De Andrade, J. P. Performance of ChatGPT on American Board of Surgery In-Training Examination preparation questions. *J. Surg. Res.* **299**, 329–335. 10.1016/j.jss.2024.04.060 (2024).
9. Gilson, A. 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.* **9**, e45312. 10.2196/45312 (2023).
10. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on medical challenge problems. arXiv. (2023). 10.48550/arXiv.2303.13375
11. Weng, T-L. et al. ChatGPT failed Taiwan's Family Medicine Board Exam. *J. Chin. Med. Association*. **86** (8), 762–766. 10.1097/JCMA.0000000000000946 (2023).
12. Zong, H. et al. Large language models in worldwide medical exams: Platform development and comprehensive analysis. *J. Med. Internet Res.* **26**, e66114. 10.2196/66114 (2024).
13. Liu, M. et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: Systematic review and meta-analysis. *J. Med. Internet Res.* **26**, e60807. 10.2196/60807 (2024).
14. Jin, Y. et al. Better to ask in English: Cross-lingual evaluation of large language models for healthcare queries. Proceedings of the ACM Web Conference. 2024:2627–2638. (2024). 10.1145/3589334.3645643
15. Sheikhalishahi, S. et al. Benchmarking large language models on Persian surgical subspecialty board examinations: A comparative study of ChatGPT-4o, ChatGPT-5, and Gemini 2.5 Flash. *Sci. Rep.* **16**, 21139. 10.1038/s41598-026-51934-9 (2026).
16. Brunicardi, F. C. et al. (eds) *Schwartz's Principles of Surgery* 11th edn (McGraw-Hill Education, 2019).
17. Townsend, C. M. Jr, Beauchamp, R. D., Evers, B. M. & Mattox, K. L. (eds) *Sabiston Textbook of Surgery: The Biological Basis of Modern Surgical Practice* 21st edn (Elsevier, 2022). 10.1016/C2017-0-04312-4
18. Zinner, M. J., Ashley, S. W. & Hines, O. J. (eds) *Maingot's Abdominal Operations* 13th edn (McGraw-Hill Education, 2019).
19. Ellison, E. C. et al. (eds) *Zollinger's Atlas of Surgical Operations* 11th edn (McGraw-Hill Education, 2021).
20. Park, S. H. et al. Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM): 2025 Updates. *Korean J. Radiol.* **26** (12), 1123–1132. 10.3348/kjr.2025.1522 (2025).
21. Gallifant, J. et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **31**, 60–69. 10.1038/s41591-024-03425-5 (2025).

22. Tripathi, S. et al. Development, Evaluation, and Assessment of Large Language Models (DEAL) Checklist: A Technical Report. NEJM AI. ;2(6):AIp2401106. (2025). 10.1056/AIp2401106

Appendix

Appendix 1. Prompt Used for Model Evaluation

Common prompt used in both evaluation workflows: original Persian prompt (verbatim)

[illegible]

[illegible]

English translation (not supplied to the models)

I am a university professor and a general-surgery specialist. The general-surgery promotion examination questions are provided in the attached file. State the correct answer as a single option. If you have a question, ask me.

The question-file-only workflow included the annual examination file alone. The reference-package workflow included the identical annual examination file together with the surgical document package described in the Methods section.

Figures

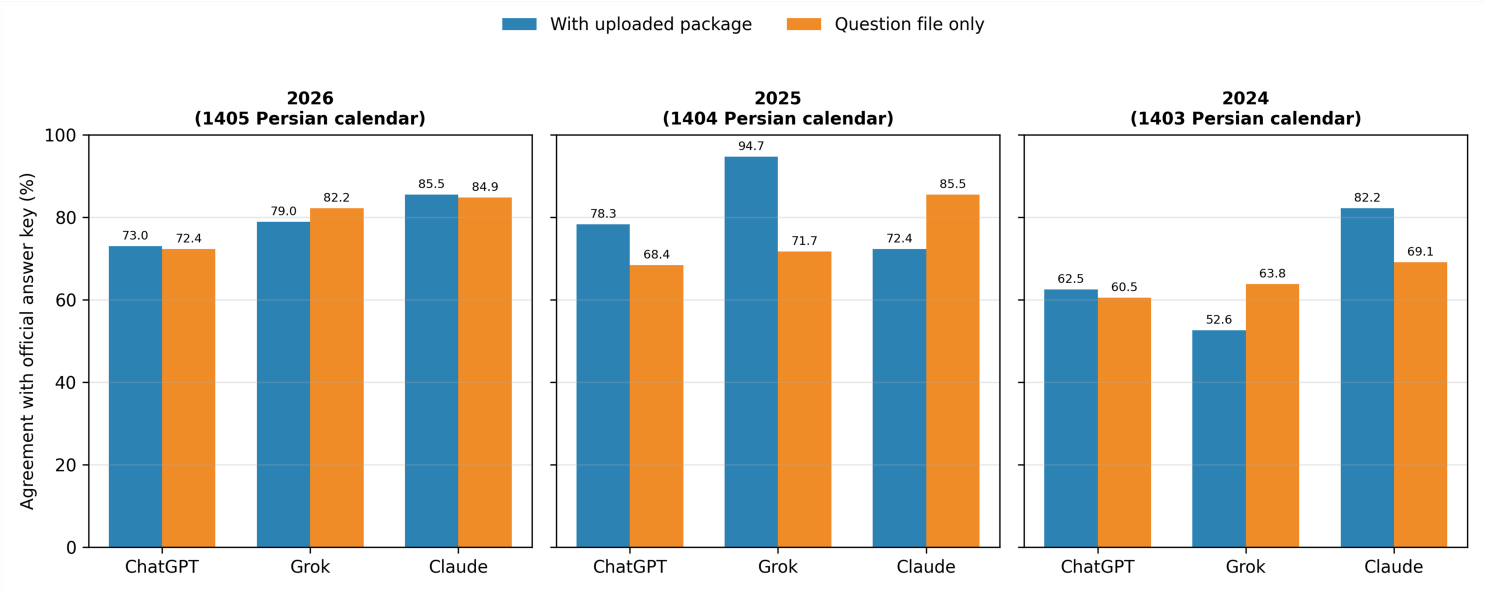


Figure 1

Agreement with the official answer key by examination year and document-upload workflow

Bars show agreement with the official answer key among 152 questions in each examination year. Examination years are shown as Gregorian year (Persian calendar year). Blue bars indicate the workflow with the uploaded surgical document package; orange bars indicate the question-file-only workflow.