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.

Evaluation of GPT-5.6 Models on the 2025 Chinese Licensed Pharmacist Qualification Examination, Including Image-Containing Questions

Jie Ji
Shantou University
Laiwen Lv
Joint Shantou International Eye Center
Liling Lin
Sun Yat-sen Memorial Hospital, Sun Yat-sen University
Jianwei Lin
ljw@jsiec.org

Joint Shantou International Eye Center

Research Article

Keywords: Chinese Licensed Pharmacist Qualification Examination, medical education, large language models, GPT-5.6, multimodal large language models

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10783773/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: To evaluate the performance of GPT-5.6-Sol, GPT-5.6-Terra, and GPT-5.6-Luna on the 2025 Chinese Licensed Pharmacist Qualification Examination (CLPQE), focusing on image-containing questions and the effect of providing image inputs.

Methods: Questions, answer keys, and images from all four CLPQE subjects were extracted from a published examination study guide. Model performance was assessed at the question and subject levels. Image-containing questions were evaluated with and without the original image files. Accuracy and 95% confidence intervals (CIs), calculated from 1,000 bootstrap replications, were used as performance measures. Overall performance was compared using Cochran's Q test and pairwise McNemar tests. Within-model comparisons of text-only and image-containing questions used chi-square tests, while the effect of providing images was assessed using McNemar tests. Bonferroni correction was applied for multiple comparisons.

Results: The dataset contained 100 questions per subject. Twenty-one image-containing questions appeared exclusively in Pharmaceutical Professional Knowledge (I). Among them, 14 had images in the stem, 9 had images in the answer options, and 2 of them had images in both the stem and the answer options. All models exceeded the passing threshold in every subject. Overall accuracies were 91.5% for GPT-5.6-Sol, 89.0% for GPT-5.6-Terra, and 83.3% for GPT-5.6-Luna, with a significant difference among models ($p < 0.0001$). GPT-5.6-Sol and GPT-5.6-Terra both outperformed GPT-5.6-Luna ($p < 0.01$). Accuracy was numerically higher for text-only questions than for image-containing questions, although these groups differed in content and potentially in difficulty. After Bonferroni correction, this difference was significant only for GPT-5.6-Sol. For image-containing questions, accuracy was numerically higher when original images were provided than when only the text and image placeholders were supplied.

Conclusion:

All three GPT-5.6 models passed every CLPQE subject, with GPT-5.6-Sol and GPT-5.6-Terra significantly outperforming GPT-5.6-Luna. Performance was generally lower on image-containing questions, but differences in question content and difficulty limit direct comparisons. Larger studies are needed to clarify the effects of image inputs. After further validation, these models may support pharmacist competency assessment, examination training, and personalized pharmaceutical education.

Background

Over the past few decades, domain-specific artificial intelligence—especially deep learning—has made great advances in medical research (1–11). More recently, general-purpose large language models (LLMs) have developed rapidly and shown increasingly strong capabilities in healthcare. A substantial number of studies have evaluated the ability of LLMs to pass medical examinations, including clinical physician qualification examinations (12–23), licensed pharmacist qualification examinations(24–30), and national nursing examinations(31). These studies have demonstrated that LLMs can achieve strong

performance on such examinations. However, most of them have relied solely on text-based data and have not incorporated multimodal information, particularly images. Because image-containing questions are often excluded, these evaluations may introduce selection bias.

Although few previous studies(21, 32) have included image-containing questions, their results, which compared performance on image-containing questions with that on text-based questions, were conflicting. Moreover, these studies have not distinguished between questions in which images appear in the question stem and those in which images appear in the answer options. Moreover, the finding that LLM performance did not decline on image-containing questions(21) warrants further investigation and substantiation.

In addition, many contemporary LLM families include multiple variants with differing computational requirements. For example, GPT-5.6 includes GPT-5.6-Sol, GPT-5.6-Terra, and GPT-5.6-Luna, while DeepSeek-V4(33) includes DeepSeek-V4-Flash and DeepSeek-V4-Pro. However, whether different variants within the same model family exhibit differences in performance on real-world medical examinations remains unclear.

To address these research gaps, this study aimed to evaluate the performance of GPT-5.6 models—including GPT-5.6-Sol, GPT-5.6-Terra, and GPT-5.6-Luna—on the 2025 Chinese Licensed Pharmacist Qualification Examination (CLPQE), with particular emphasis on image-containing questions.

The CLPQE is administered annually by China's National Medical Products Administration under the Administrative Measures for the Professional Qualification Examination for Licensed Pharmacists(34). The Chinese Licensed Pharmacist Examination is divided into two professional categories: Pharmacy and Traditional Chinese Medicine. This study evaluated only the Pharmacy category and did not assess the Traditional Chinese Medicine category. The CLPQE consists of four subjects, and the passing threshold for each subject is 60%. The four examination subjects are Pharmaceutical Professional Knowledge (I), Pharmaceutical Professional Knowledge (II), Comprehensive Pharmaceutical Knowledge and Skills, and Pharmacy Affairs Management and Regulations. Candidates must pass all subjects within a rolling four-year period.

Methods

CLPQE and Data Processing

Questions, correct answers, and images from the four subjects of the 2025 CLPQE were extracted from a published National Certified Pharmacist Qualification Examination Study Guide, a widely used CLPQE preparation book published by Henan University Press in 2026. According to the book's description and a comparison of several test-takers' recollections, the exam content in the book is identical to the real exam. The book comprises four volumes, each corresponding to a separate examination subject.

To construct the dataset, each volume of the book was first scanned and saved as a PDF file. The PDF was then converted into a Word (DOCX) document using WPS Office. After a manual review to correct character-recognition and formatting errors, a custom program was used to extract the examination identifier, question type, question stem, question options, any images, and correct answers from the DOCX file. The structured data were then saved in an Excel spreadsheet. Images were saved in JPEG format in directories named according to their question identifiers. The image files for each question were numbered sequentially. Finally, the Excel file and image files underwent a second manual review to minimize conversion errors, improve data integrity and prevent inconsistencies.

Models

The GPT-5.6 models were publicly released on July 9, 2026. According to the developer's documentation, these models demonstrated strong performance on multiple benchmark evaluations. According to OpenAI's API documentation, the knowledge cutoff for the GPT-5.6 models is February 16, 2026.

To investigate potential data leakage, we searched Baidu and Google on April 3 and May 16, 2026 did not identify a complete examination paper that could be verified as publicly available before February 16, 2026. However, these searches cannot establish the historical online availability of individual items or exclude their inclusion in proprietary training data.

Automated examination program

Models were evaluated using the Xuedingmao LLM cloud platform service(35), which is similar to OpenRouter and provides a single API endpoint for LLMs from multiple vendors. The service provider ensured that the results were consistent with those obtained by using the services directly.

A custom-designed program was developed to automatically conduct the CLPQE by leveraging the OpenAI-compatible API to communicate with the LLMs. The program used the Chat Completions API—specifically, the chat.completions.create method in the OpenAI Python library. The model identifiers used were the same as the model names. The messages sent to each LLM included a system message and a user message. Each question was evaluated independently, with no contextual carryover between questions.

The system messages used for single-choice and multiple-choice questions were originally written in Chinese and detailed in Section 1 of the supplement. Questions of Pharmacy Affairs Management and Regulations have only 4 options, so the system prompts also need to be changed accordingly. For the convenience of researchers, these system messages were translated into the following English sentences:

The system message used for single-choice questions was: *"You are a senior pharmacist. Please choose the most correct answer from the five options provided based on the question stem. Please answer directly*

with the letter of the correct option—A, B, C, D, or E—that is, you only need to provide the letter corresponding to the correct answer, without any explanation."

The system message used for multiple-choice questions was: *"You are a senior pharmacist. Please select the correct answer(s) from the five options provided based on the question stem. Each question has at least two correct answers. Please answer directly with the letters of the correct option(s)—A, B, C, D, or E—that is, you only need to provide the letters corresponding to the correct answer(s), without any explanation."*

Using the OpenAI-compatible API, the program sent text and multiple images to the LLMs. Text-based image placeholders indicated the locations of images within the text, and the image files were encoded in Base64 format. For image-containing questions evaluated without image input, the program sent only the text-based image placeholders. To reduce output variability, the temperature parameter was set to 0. Other parameters including top_p and max_tokens were set to their default values.

Parsing LLM results

A rule-based program using multiple regular expressions was used to parse LLM outputs into answer options. Four outputs could not be parsed because the model responded with “Unable to determine”; these outputs were treated as incorrect answers. For multiple-choice questions, exact-set matching was used: an answer was considered correct only if it included all the correct choices and no incorrect ones.

Statistical Analysis

The numbers and proportions of image-containing questions with images in the stem, answer options, or both locations were calculated. Accuracy and 95% confidence intervals (CIs), estimated using 1,000 bootstrap replications, were used as performance metrics (36). Cochran's Q test(37) was used to compare performance among the three models. When Cochran’s Q test indicated a significant overall difference among the models, pairwise comparisons were conducted using McNemar’s tests. Within each model, accuracy was compared between text-only and image-containing questions using the chi-square test (38). McNemar’s test(39, 40) was used to compare performance on image-containing questions when both text and images were provided versus when only text and image placeholder text were provided. The Bonferroni correction(41), applied separately within each family of tests, was used to adjust the P values from the chi-square test and McNemar’s test for multiple comparisons(42).

Two-sided exact P values were used for McNemar’s test. Cochran’s Q test used the asymptotic chi-square reference distribution. P values less than 0.05 after correction were considered statistically significant.

Application Development of the Mock Examination System

Based on our previous experience, developing and releasing an application can provide other researchers with a direct way to use this system, as well as share the data and results of this study (43).

In this study, to facilitate users' browsing of the exam questions used in this study, we developed and publicly released a mock examination system for the 2025 CLPQE supporting both English and Chinese. Anyone can register, use the actual 2025 examination data for online evaluation, and automatically receive evaluation scores and results.

The system was developed using an AI-powered development platform. Technical details about system development are available in Section 3 of the supplement.

The English version of the questions was translated from Chinese through a human-AI collaboration. Technical details about the translation process are available in Section 4 of the supplement. The translations are intended for demonstration purposes only and were not used for model evaluation. The system prompts and user prompts sent to the large language model are in Chinese.

Results

Dataset and Image Data Distribution

A dataset comprising 400 questions, with 100 questions from each examination subject, was constructed. Image-containing questions were identified only in the Pharmaceutical Professional Knowledge (I) examination. Among all 100 questions in Pharmaceutical Professional Knowledge (I), 14 (14%) had images in the stem, 9 (9%) had images in the options, and 2 (2%) had images in both, yielding 21 questions (21%) with at least one image. Examples of a text-only question, a question with an image in the stem, and a question with images in the answer options are provided in Section 2 of the supplement. One question was randomly selected for each type. In addition, all questions, images, and correct answers can be viewed using the online mock examination system.

Mock Examination System

The mock examination system is available at: http://wecom.stu.edu.cn/clpqe/. Everyone can register an account, browse the questions and images used in this study, take the online assessment, and view the correct answers.

Overall Accuracy

After aggregating the results across all four examination subjects, the overall accuracies and corresponding 95% CIs for GPT-5.6-Sol, GPT-5.6-Terra, and GPT-5.6-Luna were 91.5% (88.75%–94.25%), 89.0% (85.75%–91.75%), and 83.3% (79.75%–86.75%), respectively. Cochran’s Q test indicated a significant difference among models, $Q = 29.12$, $p < 0.0001$. The Bonferroni-adjusted P values from pairwise McNemar tests for GPT-5.6-Sol versus GPT-5.6-Terra, GPT-5.6-Sol versus GPT-5.6-Luna, and GPT-5.6-Terra versus GPT-5.6-Luna were 0.12, $< 2 \times 10^{-6}$, and 0.001, respectively. Thus, GPT-5.6-Sol and GPT-5.6-Terra each performed significantly better than GPT-5.6-Luna, whereas the difference between GPT-5.6-Sol and GPT-5.6-Terra was not statistically significant.

Subject-level Accuracy

Figure 1 shows the accuracies and 95% CIs of the GPT-5.6 models across all subjects in the 2025 CLPQE. On the x-axis, Exam I, Exam II, Comprehensive, and Regulations represent Pharmaceutical Professional Knowledge (I), Pharmaceutical Professional Knowledge (II), Comprehensive Pharmaceutical Knowledge and Skills, and Pharmacy Affairs Management and Regulations, respectively. The red dotted line indicates the passing score.

Table 1 presents subject-level accuracy, 95% CIs, and pass/fail results for the GPT-5.6 models across the four subjects of the 2025 CLPQE.

Table 1
Subject-level accuracy, 95% CIs, and pass/fail results

Examination	Model	Correct / Total	Accuracy (95% CI)	Passed / Failed
Exam I	GPT-5.6-Sol	93 / 100	0.93 (0.87–0.98)	Passed
Exam I	GPT-5.6-Terra	86 / 100	0.86 (0.79–0.92)	Passed
Exam I	GPT-5.6-Luna	87 / 100	0.87 (0.80–0.93)	Passed
Exam II	GPT-5.6-Sol	94 / 100	0.94 (0.89–0.98)	Passed
Exam II	GPT-5.6-Terra	94 / 100	0.94 (0.89–0.98)	Passed
Exam II	GPT-5.6-Luna	88 / 100	0.88 (0.81–0.94)	Passed
Comprehensive	GPT-5.6-Sol	96 / 100	0.96 (0.92–0.99)	Passed
Comprehensive	GPT-5.6-Terra	96 / 100	0.96 (0.92–0.99)	Passed
Comprehensive	GPT-5.6-Luna	91 / 100	0.91 (0.84–0.96)	Passed
Regulations	GPT-5.6-Sol	83 / 100	0.83 (0.75–0.90)	Passed
Regulations	GPT-5.6-Terra	80 / 100	0.80 (0.72–0.88)	Passed
Regulations	GPT-5.6-Luna	67 / 100	0.67 (0.57–0.76)	Passed

Accuracies of Text-Only and Image-Containing Questions

Table 2 presents model performance on text-only and image-containing questions in Pharmaceutical Professional Knowledge (I), along with the results of statistical tests. For image-containing questions, both the text and the image were provided to the LLMs. Each entry in columns 2 and 3 reports the number of correct answers, the total number of questions, and the resulting accuracy. Table 2 illustrates that for all models, the accuracy for text-only questions was numerically higher than that for image-containing questions. The accuracy differences between image-containing and text-only questions were − 15.25, − 6.39, and − 13.68 percentage points for GPT-5.6-Sol, GPT-5.6-Terra, and GPT-5.6-Luna, respectively. The difference was statistically significant for GPT-5.6-Sol ($p < 0.05$). Bonferroni-adjusted p-values were calculated by multiplying raw p-values by the number of comparisons and truncating values at 1.00.

Table 2
Accuracy and statistical test results of text-only questions and image-containing questions in Exam I

Model name	Text-only questions	Image-containing questions	Difference in accuracy (image-containing minus text-only)	P value	P value after Bonferroni correction
GPT-5.6-Sol	76/79 (96.2%)	17/21 (81.0%)	-15.25%	0.015	0.045
GPT-5.6-Terra	69/79 (87.3%)	17/21 (81.0%)	-6.39%	0.45	1.000
GPT-5.6-Luna	71/79 (89.9%)	16/21 (76.2%)	-13.68%	0.098	0.29

Table 3 reports the accuracies on all image-containing questions, questions with images only in the stem, and questions with images only in the options in Pharmaceutical Professional Knowledge (I), both with and without transmitting the image itself. Each cell in columns 3, 4, and 5 reports the number of correct answers, the total number of questions, and the corresponding accuracy. The number of questions with images in the stem or the options is fewer than the sum of the number of questions with images in the stem and the number with images in the options, because there are 2 questions that contain images in both the stem and the options.

Table 3
Accuracy on image-containing questions in Exam I with and without sending actual image files

Model name	Image input	Images in the stem or the options	Images in only the stem	Images in only the options
GPT-5.6-Sol	Provided	17/21 (81.0%)	10/12 (83.3%)	5/7 (71.4%)
GPT-5.6-Sol	Not provided	14/21 (66.7%)	10/12 (83.3%)	4/7 (57.1%)
GPT-5.6-Terra	Provided	17/21 (81.0%)	10/12 (83.3%)	5/7 (71.4%)
GPT-5.6-Terra	Not provided	13/21 (61.9%)	8/12 (66.7%)	3/7 (42.9%)
GPT-5.6-Luna	Provided	16/21 (76.2%)	9/12 (75.0%)	5/7 (71.4%)
GPT-5.6-Luna	Not provided	12/21 (57.1%)	8/12 (66.7%)	3/7 (42.9%)

Table 3 shows that for GPT-5.6-Sol, GPT-5.6-Terra, and GPT-5.6-Luna, providing images for image-containing questions (regardless of whether images appeared in the question stem or options) improved accuracy by 14.3, 19.0, and 19.0 percentage points compared with not providing the images,

respectively. However, the Bonferroni-adjusted P values from paired McNemar tests were 0.75, 0.38, and 0.66 for GPT-5.6-Sol, GPT-5.6-Terra, and GPT-5.6-Luna, respectively. None of these differences was statistically significant, possibly because of the small sample size of image-containing questions. In addition, because the stem and option subgroups overlapped and were small, subgroup differences were exploratory and were not formally tested.

Discussion

All evaluated models achieved observed accuracies above the predefined 60% passing threshold in every subject, often by substantial margins. Overall, GPT-5.6-Sol and GPT-5.6-Terra each performed significantly better than GPT-5.6-Luna, whereas the difference between GPT-5.6-Sol and GPT-5.6-Terra was not statistically significant.

Text-only questions in Pharmaceutical Professional Knowledge I had higher observed accuracy than image-containing questions. These findings differ from those reported by a previous study (21). For questions containing images, accuracy was numerically higher when the question text and actual images were provided than when only the question text and placeholder text for the images were provided, particularly when the images appeared in the answer options. One possible explanation is that images in question stems may sometimes be illustrative rather than essential for answering the question. In contrast, images presented as answer options may be necessary for selecting the correct response.

This study has several notable strengths. To our knowledge, this study is among the first to evaluate LLM performance on a pharmacist qualification examination both with and without sending actual image files and to distinguish between images in question stems and images in answer options. Another strength is the evaluation of three GPT-5.6 model variants on the most recent CLPQE available at the time of the study.

Despite these promising results, several limitations should be acknowledged. First, the knowledge cutoff date for the GPT-5.6 models was February 16, 2026, which was later than the administration date of the 2025 CLPQE. Although the guidebook was published after the knowledge cutoff date, we searched commonly used internet search engines, including Baidu and Google, and did not identify comprehensive online content related to the 2025 CLPQE. Nevertheless, the possibility that some examination content was available online or was otherwise included in the models' training data cannot be excluded.

Second, although numerical differences in accuracy between text-only and image-containing questions were observed for all models, only the comparison for GPT-5.6-Sol was statistically significant.

Third, none of the McNemar tests comparing LLM accuracy on image-containing questions when actual image files were provided versus when they were not were statistically significant. In addition, although the observed improvement associated with image provision was numerically larger for questions containing images in the answer options than for those containing images in the stem, this study does

not provide reliable evidence for this difference. The subgroup samples were small and overlapping, and no formal comparison between subgroups was conducted.

Fourth, the limited sample size reduced the precision and generalizability of the findings. The evaluation was restricted to the 2025 CLPQE, and earlier examinations were excluded to minimize the risk of data leakage. Fifth, text-only questions had higher observed accuracy than image-containing questions. Because these question sets may differ in difficulty and content, this difference should not be interpreted solely as evidence of weaker multimodal capability. Further research is needed to determine the reasons underlying this difference.

Sixth, we evaluated only models in the GPT-5.6 series and did not include models from other model families, such as Qwen(44), DeepSeek(33), or Claude.

The clinical translational value of this study lies in the application of LLMs to pharmacy education and can be implemented in two major settings. First, it can serve licensed pharmacist examination training: an intelligent test-preparation tutoring system can be developed to generate customized practice plans targeting individual candidates' knowledge gaps. In addition, it can be applied to competency assessment and gap-filling for pharmaceutical professionals: the evaluation framework can be adapted into an assessment tool for pharmacist job competency, quickly identifying knowledge weaknesses of medicine professionals in modules such as basic pharmaceutical theory, identification, and clinical medication use, while providing corresponding learning resources and individualized continuing education. Future research may further enrich the assessment question types and promote the transformation of this evaluation system from research settings to educational training and industry management contexts.

Conclusion

All evaluated models achieved passing scores on the 2025 CLPQE, with substantial margins. GPT-5.6-Sol and GPT-5.6-Terra performed significantly better than GPT-5.6-Luna. Observed accuracy was higher for text-only questions than for image-containing questions; however, these question groups may have differed in content and difficulty. Larger studies are needed because several observed differences did not reach statistical significance. After thorough validation, these models are expected to hold practical value in the future in areas such as competency assessment of pharmaceutical professionals, training for licensed pharmacist examinations, and personalized learning for pharmaceutical professionals.

Abbreviations

LLM: large language model

CLPQE: Chinese Licensed Pharmacist Qualification Examination

CI: confidence interval

Declarations

Ethics approval and consent to participate

Institutional review board approval was not required because this study used only publicly available examination-preparation materials and did not involve human participants or identifiable personal data.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2026A1515012879 to Liling Lin) and the Guangdong Provincial Research Project on Teaching Reform for Clinical Teaching Bases (Grant No. 2025JD148).

Author Contribution

Conceptualization, literature searching, and performance analysis: Jie Ji, Liling Lin, and Jianwei Lin. Writing original draft: Jie Ji, and Laiwen Lv. Data processing: Jie Ji and Laiwen Lv. Visualization: Jie Ji, Laiwen Lv, Liling Lin and Jianwei Lin. Review: Liling Lin and Jianwei Lin. Supervision and validation: Liling Lin and Jianwei Lin.

Acknowledgements

Not applicable.

Data Availability

The dataset can be viewed online at http://wecom.stu.edu.cn/clpqe/, a system developed specifically for this study. The code and results are publicly available at https://github.com/linchundan88/clpqe.

References

1. Cen L-P, Ji J, Lin J-W, Ju S-T, Lin H-J, Li T-P, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. Nature Communications. 2021;12(1).
2. Wang J, Ji J, Zhang M, Lin J-W, Zhang G, Gong W, et al. Automated Explainable Multidimensional Deep Learning Platform of Retinal Images for Retinopathy of Prematurity Screening. JAMA network open. 2021;4(5):e218758-e.
3. Ji J, Zhang W, Dong Y, Lin R, Geng Y, Hong L. Automated cervical cell segmentation using deep ensemble learning. BMC Medical Imaging. 2023;23(1):137.
4. Huang J, Jie J, Ma H, Xie S, Liao H, Ouyang K, et al. Deep learning for automated mandibular canal segmentation in CBCT scans. BMC Oral Health. 2025;25(1):1699.
5. Yuan L, Kang D, Dong X, Liu L, Grzybowski A, Jin K. Artificial intelligence in clinical education in ophthalmology: a systematic review. Visual Neuroscience. 2025.
6. Shao Y, Lou Y, Gao Y, Tan G, Liu T, Yu H, et al. Expert consensus on the application of artificial intelligence in neuro-ophthalmic diseases. Visual Neuroscience. 2025;42(1):e019.
7. Zheng X, Zhang C, Lei D, Zhou Y, Tang N, Tang F, et al. Relationship between systemic indicators and retinal vessel caliber in patients with stable angina. Visual Neuroscience. 2025;42(1):e025.
8. O'Hare L, Hibbard PB. Support for the efficient coding account of visual discomfort. Visual Neuroscience. 2024;41:E008.
9. Qiu J, Lin J, Shi L, Huang C, Qiu K, Wang G. Multidimensional imaging model for diagnosing primary open-angle glaucoma using fundus photographs and optical coherence tomography images. BMJ open ophthalmology. 2025;10:e002274.
10. Zhang Z, Duan Y, Lin J, Luo W, Lin L, Gao Z. Artificial intelligence in anesthesia: insights from the 2024 Nobel Prize in Physics. Anesthesiology and Perioperative Science. 2025;3(1):5.
11. Xu M, Lin J, Chen S, Huang Y. The Application of Artificial Intelligence in Epiretinal Membrane from Detection to Prognosis Prediction. Seminars in Ophthalmology. 2026;41(5):594–603.
12. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023;9:e45312.
13. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health. 2023;2(2):e0000198.
14. Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. arXiv e-prints. 2023:arXiv:2303.18027.
15. Liu M, Okuhara T, Huang W, Ogihara A, Nagao HS, Okada H, et al. Large Language Models in Dental Licensing Examinations: Systematic Review and Meta-Analysis. International Dental Journal. 2025;75(1):213 – 22.
16. Ohta K, Ohta S. The Performance of GPT-3.5, GPT-4, and Bard on the Japanese National Dentist Examination: A Comparison Study. Cureus. 2023;15.

17. Bicknell BT, Butler D, Whalen S, Ricks J, Dixon CJ, Clark AB, et al. ChatGPT-4 Omni Performance in USMLE Disciplines and Clinical Skills: Comparative Analysis. JMIR Med Educ. 2024;10:e63430.
18. Lin S-Y, Chan PK, Hsu W-H, Kao C-H. Exploring the proficiency of ChatGPT-4: An evaluation of its performance in the Taiwan advanced medical licensing examination. DIGITAL HEALTH. 2024;10:20552076241237678.
19. Revilla-León M, Barmak AB, Sailer I, Kois J, Att W. Performance of an Artificial Intelligence-Based Chatbot (ChatGPT) Answering the European Certification in Implant Dentistry Exam. The International Journal of Prosthodontics. 2024:1–5.
20. Sismanoglu S, Capan BS. Performance of artificial intelligence on Turkish dental specialization exam: can ChatGPT-4.0 and gemini advanced achieve comparable results to humans? BMC Medical Education. 2025;25(1):214.
21. Miyazaki Y, Hata M, Omori H, Hirashima A, Nakagawa Y, Eto M, et al. Performance of ChatGPT-4o on the Japanese Medical Licensing Examination: Evaluation of Accuracy in Text-Only and Image-Based Questions. JMIR Med Educ. 2024;10:e63129.
22. Yaneva V, Baldwin P, Jurich DP, Swygert K, Clauser BE. Examining ChatGPT Performance on USMLE Sample Items and Implications for Assessment. Academic Medicine. 2024;99(2):192-7.
23. Diao Y, Wu M, Xu J, Pan Y. Multiple Large Language Models' Performance on the Chinese Medical Licensing Examination: Quantitative Comparative Study. JMIR Hum Factors. 2025;12:e77978.
24. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Medical Education. 2024;24(1).
25. Ehlert A, Ehlert B, Cao B, Morbitzer K. Large Language Models and the North American Pharmacist Licensure Examination (NAPLEX) Practice Questions. American Journal of Pharmaceutical Education. 2024;88(11):101294.
26. Chen T-J, Wang Y-M, Shen H-W. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. Journal of the Chinese Medical Association. 2023;86(7):653-8.
27. Kunitsu Y. The Potential of GPT-4 as a Support Tool for Pharmacists: Analytical Study Using the Japanese National Examination for Pharmacists. JMIR Med Educ. 2023;9:e48452.
28. Jin HK, Kim E. Performance of GPT-3.5 and GPT-4 on the Korean Pharmacist Licensing Examination: Comparison Study. JMIR Med Educ. 2024;10:e57451.
29. Huang C, Sun Y, Liu W. A comparative study of the performance of different large language models in the Chinese National Pharmacist Licensing Examination. Front Med (Lausanne). 2026;Volume 13–2026.
30. Sato H, Ogasawara K, Sakurai H. Performance Evaluation of 18 Generative AI Models (ChatGPT, Gemini, Claude, and Perplexity) in 2024 Japanese Pharmacist Licensing Examination: Comparative Study. JMIR Med Educ. 2025;11:e76925.

31. Takahashi Y, Kumakura R, Okamoto R, Omote S. Performance of Large Language Models in the Japanese Public Health Nurse National Examination: Comparative Cross-Sectional Study. JMIR Nursing. 2026;9:e82842.
32. Safavi Naini SAA, Ali S, Shahab O, Shahhoseini Z, Savage T, Rafiee S, et al. Vision-Language and Large Language Model Performance in Gastroenterology: GPT, Claude, Llama, Phi, Mistral, Gemma, and Quantized Models2024.
33. DeepSeek AI, Xu A, Lin B, Xue B, Wang B, Xu B, et al. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence. arXiv e-prints. 2026:arXiv:2606.19348.
34. Administration CNMP. Administrative Measures for the Professional Qualification Examination for Licensed Pharmacists 2019.
35. Xuedingmao large language models cloud platform.
36. Thomas JD, Bradley E. Bootstrap confidence intervals. Statistical Science. 1996;11(3):189–228.
37. Cochran WG. The Comparison of Percentages in Matched Samples. Biometrika. 1950;37(3/4):256 – 66.
38. McHugh ML. The chi-square test of independence. Biochem Med (Zagreb). 2013;23(2):143-9.
39. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153-7.
40. Dietterich TG. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural computation. 1998;10(7):1895 – 923.
41. Zheng C, Zeng R, Wu G, Hu Y, Yu H. Investigating causal associations between the genetic liability to primary open-angle glaucoma and psychiatric disorders. Visual Neuroscience. 2025;42(1):e017.
42. Armstrong RA. When to use the Bonferroni correction. Ophthalmic and Physiological Optics. 2014;34(5):502-8.
43. Ou Y, Liao X, Zhu Z, Zhang D, Chen Y, Yu S, et al. Monitoring and improvement of posture compliance in patients with intraocular tamponade. Visual Neuroscience. 2025;42(1):e014.
44. Yang A, Yang B, Hui B, Zheng B, Yu B, Zhou C, et al. Qwen2 Technical Report. arXiv e-prints. 2024:arXiv:2407.10671.

Figures

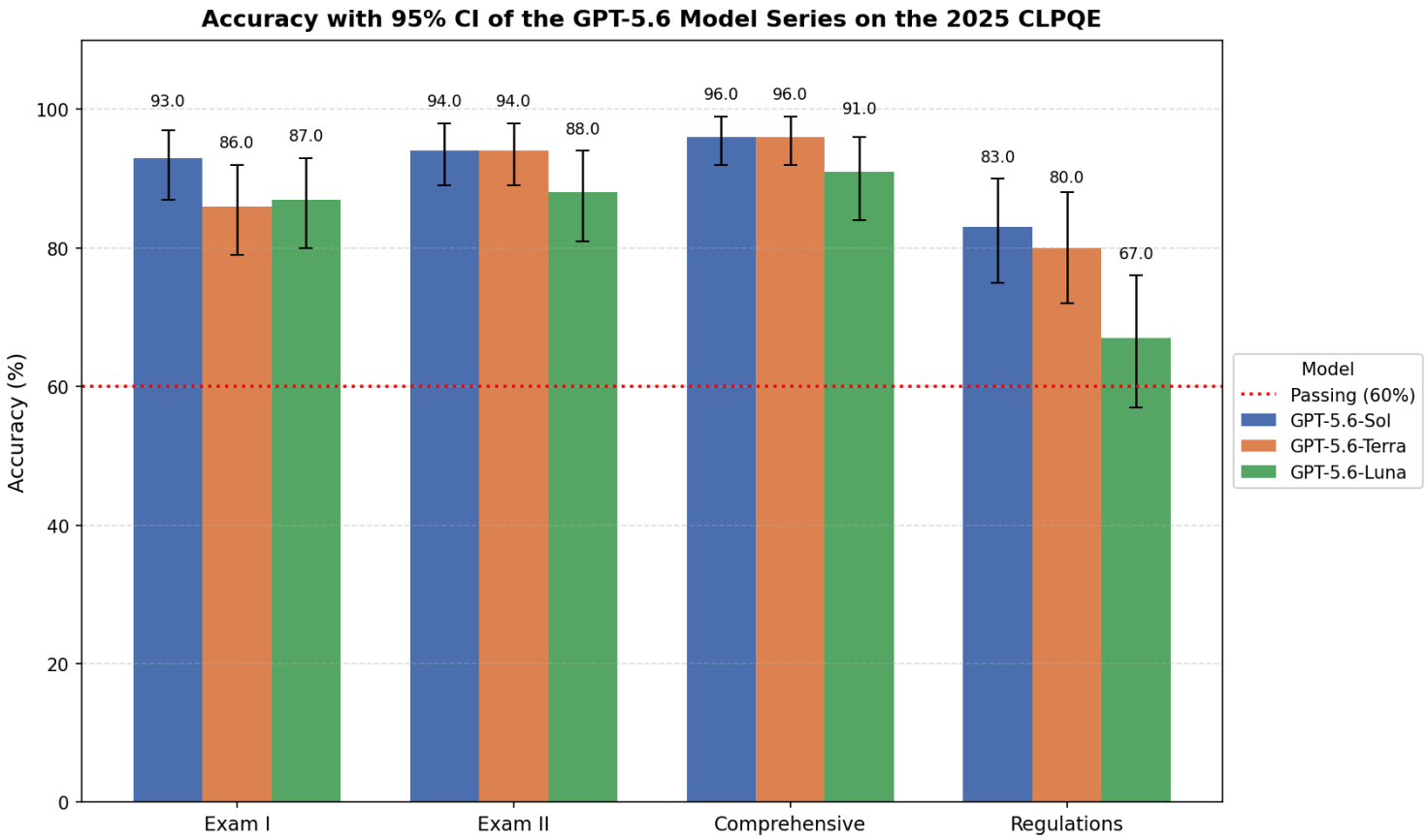


Figure 1

Accuracy with 95% confidence intervals for the GPT-5.6 models across all subjects of the 2025 CLPQE.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement2026822.docx