

Original Paper

Qwen-2.5 Outperforms Other Large Language Models in the Chinese National Nursing Licensing Examination: Retrospective Cross-Sectional Comparative Study

Shiben Zhu^{1,2}, MSc; Wanqin Hu³, MNS; Zhi Yang³, MNS; Jiani Yan⁴, MIM; Fang Zhang⁴, BBA

¹Department of Infectious Diseases, Nanfang Hospital, Southern Medical University, Guangzhou, China

²State Key Laboratory of Organ Failure Research, Key Laboratory of Infectious Diseases Research in South China, Ministry of Education, Guangdong Provincial Key Laboratory of Viral Hepatitis Research, Guangdong Provincial Clinical Research Center for Viral Hepatitis, Guangdong Institute of Hepatology, Guangzhou, China

³School of Nursing and Health Studies, Hong Kong Metropolitan University, Kowloon, Hong Kong, China (Hong Kong)

⁴Department of Science and Education, Shenzhen Baoan Women's and Children's Hospital, Shenzhen, China

Corresponding Author:

Fang Zhang, BBA

Department of Science and Education

Shenzhen Baoan Women's and Children's Hospital

56 Yulu Road

Xin'an Street, Bao'an District

Shenzhen, 518001

China

Phone: 86 13686891225

Email: zhangfangf11@163.com

Abstract

Background: Large language models (LLMs) have been proposed as valuable tools in medical education and practice. The Chinese National Nursing Licensing Examination (CNNLE) presents unique challenges for LLMs due to its requirement for both deep domain-specific nursing knowledge and the ability to make complex clinical decisions, which differentiates it from more general medical examinations. However, their potential application in the CNNLE remains unexplored.

Objective: This study aims to evaluate the accuracy of 7 LLMs including GPT-3.5, GPT-4.0, GPT-4o, Copilot, ERNIE Bot-3.5, SPARK, and Qwen-2.5 on the CNNLE, focusing on their ability to handle domain-specific nursing knowledge and clinical decision-making. We also explore whether combining their outputs using machine learning techniques can improve their overall accuracy.

Methods: This retrospective cross-sectional study analyzed all 1200 multiple-choice questions from the CNNLE conducted between 2019 and 2023. Seven LLMs were evaluated on these multiple-choice questions, and 9 machine learning models, including Logistic Regression, Support Vector Machine, Multilayer Perceptron, k-nearest neighbors, Random Forest, LightGBM, AdaBoost, XGBoost, and CatBoost, were used to optimize overall performance through ensemble techniques.

Results: Qwen-2.5 achieved the highest overall accuracy of 88.9%, followed by GPT-4o (80.7%), ERNIE Bot-3.5 (78.1%), GPT-4.0 (70.3%), SPARK (65.0%), and GPT-3.5 (49.5%). Qwen-2.5 demonstrated superior accuracy in the Practical Skills section compared with the Professional Practice section across most years. It also performed well in brief clinical case summaries and questions involving shared clinical scenarios. When the outputs of the 7 LLMs were combined using 9 machine learning models, XGBoost yielded the best performance, increasing accuracy to 90.8%. XGBoost also achieved an area under the curve of 0.961, sensitivity of 0.905, specificity of 0.978, F_1 -score of 0.901, positive predictive value of 0.901, and negative predictive value of 0.977.

Conclusions: This study is the first to evaluate the performance of 7 LLMs on the CNNLE and that the integration of models via machine learning significantly boosted accuracy, reaching 90.8%. These findings demonstrate the transformative potential of LLMs in revolutionizing health care education and call for further research to refine their capabilities and expand their impact on examination preparation and professional training.

(*JMIR Med Inform* 2025;13:e63731) doi: [10.2196/63731](https://doi.org/10.2196/63731)

KEYWORDS

large language models; LLMs; Chinese National Nursing Licensing Examination; ChatGPT; Qwen-2.5; multiple-choice questions;

Introduction

Nursing licensure examinations are essential for maintaining professional standards, ensuring that health care systems are staffed with qualified professionals, and safeguarding patient safety [1]. These examinations assess nurses' clinical judgment, decision-making, and practical skills, ensuring high-quality care and fostering public trust in the profession [2]. Upholding rigorous standards is critical, as competent health care professionals are crucial for addressing the diverse and complex needs of patients worldwide. The Chinese National Nursing Licensing Examination (CNNLE) plays an important role in maintaining high standards of nursing care in China, ensuring that graduates are well prepared for professional practice [3]. Serving as a benchmark for nursing competence, the CNNLE confirms that nurses possess the necessary skills and knowledge to provide safe and effective care [4]. Beyond its impact on health care quality, the CNNLE also influences educational policies, guiding nursing curricula to meet evolving health care demands. As health care becomes more complex, innovative tools are needed to support the development of skilled professionals capable of providing effective patient care.

The integration of artificial intelligence (AI) in education is transforming learning and assessment, particularly in fields such as nursing [5-7]. ChatGPT, an AI tool that generates content by identifying patterns in its training data, simulates human-like conversations and answers questions across a wide range of topics [8,9]. Its ability to provide correct answers and offer immediate, detailed feedback makes it a valuable resource for students in simulated test environments and question banks [10]. This success in examination settings has sparked interest in using ChatGPT as a self-learning tool, suggesting its potential for enhancing examination preparation and knowledge development [11]. Large language models (LLMs) hold promise for clinical education [12], where these models integrate natural language processing with user-friendly interfaces [13]. In clinics, LLMs are increasingly valuable, particularly in diagnosis [14,15] and clinical licensing examinations [16], where accuracy is crucial. Tools such as ChatGPT are being recognized for their potential to enhance clinical documentation [17], improve diagnostic accuracy [18], and streamline patient care workflows [19]. However, the rapid development of LLMs presents significant challenges in assessing their reliability in the CNNLE.

Passing CNNLE demands not only theoretical knowledge but also clinical decision-making, critical thinking, and practical skills, areas where LLMs often underperform [9]. While tools such as ChatGPT have demonstrated an overall accuracy of 80.75% in nursing education [20], their effectiveness diminishes with complex, context-specific questions requiring nuanced medical knowledge [21-23]. Moreover, concerns regarding patient privacy [24] and biases [25] in LLM outputs raise questions about their suitability for high-stakes assessments such as the CNNLE, which emphasize fairness and accuracy [26,27]. Despite the growing interest in LLMs for medical

education, their potential in the CNNLE remains unexplored. Limited understanding exists regarding their ability to handle clinical reasoning, contextual interpretation, and multistep problem-solving in this specific setting. Addressing this gap is crucial to assess their reliability, limitations, and transformative potential in clinical education. Here, this study examines the distribution of question types in the CNNLE from 2019 to 2023 and evaluates the accuracy of 7 LLMs including GPT-3.5, GPT-4.0, GPT-4o, Copilot, ERNIE Bot-3.5, SPARK, and Qwen-2.5—in addressing domain-specific nursing knowledge and clinical decision-making. Furthermore, the study explores whether combining their outputs through machine learning techniques can enhance overall accuracy in this context.

Methods

Study Design

This retrospective cross-sectional study evaluated the performance of 7 LLMs on 1200 multiple-choice questions (MCQs) from the CNNLE administered between 2019 and 2023. The study design was chosen for its suitability in systematically analyzing preexisting datasets and providing the capabilities of LLMs across various question types and levels of complexity. A head-to-head evaluation approach was adopted to compare the LLMs. Each MCQ was independently input into each model under identical conditions, ensuring consistency and fairness in the assessment. This parallel evaluation minimized variability caused by external factors, such as differences in question formats or content, allowing for a direct comparison of performance across all models. By using historical data and using a head-to-head evaluation, this study provides an analysis of LLM performance in nursing licensure examinations, offering their potential applications in nursing education and assessment.

Data Collection

This study analyzed all 1200 MCQs from the CNNLE administered between 2019 and 2023. Each year, 240 MCQs were included, encompassing the 4 question types (A1, A2, A3, and A4) although their proportions varied annually. This comprehensive approach ensured that the evaluation covered diverse question formats and varying levels of complexity, reflecting the full scope of the CNNLE. To ensure the integrity of the evaluation process, 2 researchers (SZ and WH) independently entered each question into 7 LLMs on separate computers. Each question was input into a new chat session to prevent any influence from prior interactions. The LLMs generated answers and explanations solely based on the input questions without pretraining instructions or additional prompts.

If inconsistencies were detected in the responses, a third computer was used to reenter the question in a fresh chat session after clearing the LLMs' memory. In such cases, the models were instructed to provide more detailed explanations. The researchers then collectively reviewed the answers and explanations to determine the most accurate and contextually appropriate response. When LLMs exhibited confusion, failed

to provide explanations, produced multiple answers including the correct one, or encountered specialized queries (eg, questions on local policies), additional instructions were provided. These instructions included prompts such as, “This is a single-choice question. Please select the most suitable or probable answer from options 1 to 5,” “Please choose the incorrect option,” “Tell me the reason why,” “In Chinese local policy,” “In Chinese local law,” and “In Chinese society.” All data generated or analyzed during this study are provided in [Multimedia Appendix 1](#), and the iPython Jupyter notebook code is available in [Multimedia Appendix 2](#).

Ethical Considerations

The evaluations were conducted between May 15, 2024, and July 17, 2024. All responses were cross-verified against the official CNNLE answer keys. These measures enhanced the reliability and validity of the evaluation process. As this study was purely analytical and did not involve human participants, institutional review board approval and informed consent were not required. All collected data were fully anonymized by removing names, contact details, and other direct identifiers, ensuring no means to reidentify participants.

Measurements

The CNNLE

The CNNLE [28] comprises 2 sections: Professional Practice and Practical Skills, each with 120 questions per unit. The Professional Practice section evaluates a candidate’s ability to implement nursing-related knowledge in clinical settings in a safe and effective manner. It covers medical knowledge related to health and disease, basic nursing skills, and the application of social and humanistic knowledge in nursing practice. The Practical Skills section assesses candidates’ capability to apply nursing knowledge and skills in performing nursing tasks. Topics include clinical manifestations of diseases, treatment principles, health assessment, nursing procedures, professional nursing techniques, and health education. The examination format involves objective questions presented in a computer-based format.

The examination includes 4 question types: A1, A2, A3, and A4, all of which are MCQs. A1 and A2 questions are relatively straightforward, focusing on single knowledge points and brief clinical case summaries, respectively. A3 and A4 questions involve shared clinical scenarios, requiring candidates to analyze and synthesize information comprehensively. A3 questions present 2-3 distinct, patient-centered clinical situations, while A4 questions depict more complex scenarios involving a single patient or family, with 4-6 independent questions that may introduce new information sequentially to test clinical integration skills.

LLM Selection

We selected 7 LLMs including GPT-3.5, GPT-4.0, GPT-4o, Copilot, ERNIE Bot-3.5, SPARK, and Qwen-2.5. This diverse selection enabled a comprehensive examination of LLM performance under standardized conditions. GPT-3.5, developed by OpenAI and released in March 2022, is known for generating coherent and contextually relevant text. GPT-4.0, released by

OpenAI in March 2023, offers significant improvements in accuracy and understanding. GPT-4o, introduced in May 2024, is an optimized version of GPT-4.0, designed for enhanced performance. ERNIE Bot-3.5, created by Baidu and released in June 2023, is tailored for understanding and generating text in Chinese. SPARK, developed by iFLYTEK and launched in May 2023, enhances performance tools by providing intelligent assistance. Qwen-2.5, created by Alibaba and launched in May 2024, is optimized for complex language understanding, particularly in shopping and customer support contexts. To ensure effectiveness and reliability, each inquiry was conducted only once in a new chat session with each LLM, using 2 different computers. This approach aimed to evaluate the designs’ efficiency in real-world situations without the influence of responses loopholes.

Machine Learning Models

We selected 9 machine learning models, each with recognized performance in classification tasks. Logistic Regression (LR) [29] is a fundamental linear model made use of for binary category tasks as a result of its simplicity and interpretability. Support Vector Machine (SVM) [30] excels in high-dimensional and complicated settings, giving durable classification efficiency. Multilayer Perceptron (MLP) [31] is a neural network model that properly identifies complicated patterns with its split structure. The k-nearest neighbors (KNN) [32] algorithm is an uncomplicated, nonparametric supervised learning approach that categorizes or predicts information factors based upon their proximity to bordering points, extensively acknowledged for its simplicity and effectiveness in both category and regression jobs.

Ensemble models improve prediction performance by incorporating multiple designs to alleviate overfitting and improve generalization. Random Forest (RF) [33] is esteemed for its high precision and ability to alleviate overfitting with ensemble learning, accumulating the predictions of choice trees using a majority ballot to enhance anticipating toughness. Light Gradient-Boosting Machine (LightGBM) [34] is a highly reliable gradient increasing framework that makes use of a histogram-based technique to bin constant features, speeding up training speed, enhancing memory usage, and mastering processing massive datasets with impressive rate and effectiveness. Adaptive Boosting (AdaBoost) [35] prioritizes tough situations, enhancing category precision by iteratively changing weights to boost the design. Extreme Gradient Boosting (XGBoost) [36], a sophisticated slope-boosting system developed by Chen, iteratively refines designs by splitting tree nodes and suitable residuals, demonstrating extraordinary scalability and superior efficiency throughout varied applications. CatBoost [37], introduced in 2018, is a sophisticated gradient boosting algorithm known for its outstanding handling of specific functions, reduced training times, and the use of a money-grubbing technique to pinpoint ideal tree divides, thereby improving forecast precision.

Statistical Analysis

The statistical analysis was conducted using Python 3.11.5 (Python Software Foundation) within the Microsoft Visual Studio Code environment. In preparing the dataset, responses

where the LLMs failed to provide any answer were categorized as missing values and coded as -1 . For valid responses labeled (A, B, C, D, and E), a numerical encoding scheme was applied, converting them to (1, 2, 3, 4, and 5), respectively. To prepare the data for machine learning algorithms, the dataset underwent normalization, scaling all features to a range between 0 and 1 using the MinMaxScaler from the Scikit-learn library. Descriptive statistics were used to analyze the distribution of question types within the CNNLE dataset from 2019 to 2023. Furthermore, accuracy percentages for the LLMs were computed across 2 distinct subjects and 4 different question types. Various machine learning models were then used with the objective of enhancing predictive performance.

Nine machine learning models, including LR, SVM, MLP, KNN, RF, LightGBM, AdaBoost, XGBoost, and CatBoost, were trained specifically for this task using the processed CNNLE dataset. None of the models were pretrained; instead, they were trained and optimized using hyperparameter tuning tailored to the dataset. For instance, parameters such as the number of trees and maximum depth were adjusted for RF, while learning rates and boosting parameters were optimized for LightGBM and XGBoost. The leave-one-out cross-validation method was used to ensure robustness and reliability. The dataset was split into training (90%) and testing (10%) sets, with the training set further divided into 9 subsets for hyperparameter tuning. This iterative process was repeated until each subset served as a validation set, minimizing overfitting and ensuring robust performance metrics for the models.

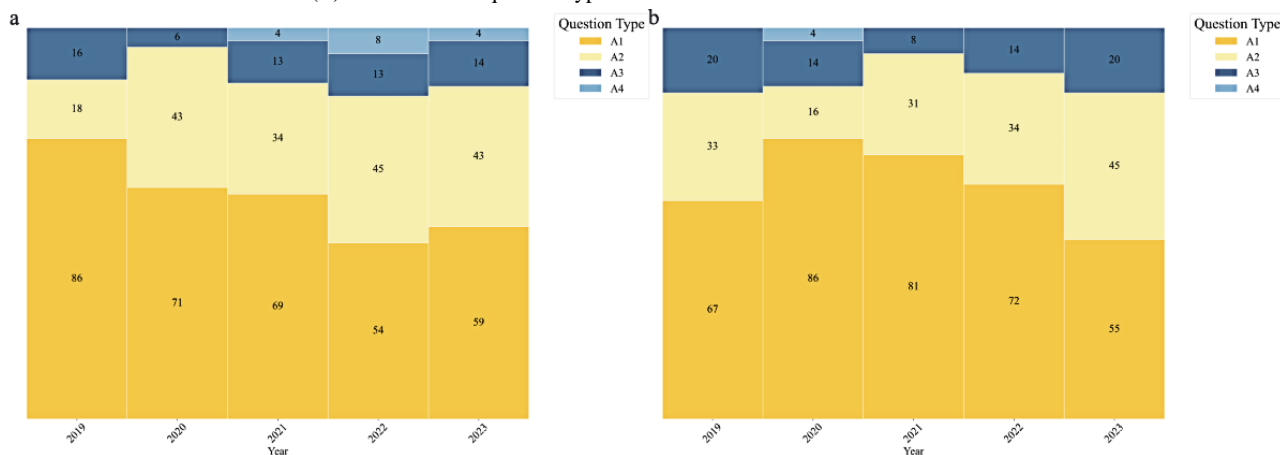
Model performance was assessed using correlation heatmaps, area under the curves (AUCs), and 7 evaluation metrics: AUC, sensitivity, specificity, F_1 -score, accuracy, positive predictive value (PPV), and negative predictive value (NPV). Feature importance was analyzed using Shapley Additive Explanations (SHAP), providing the contributions of individual features. SHAP analysis focused on understanding the relative contributions of the 7 LLMs, highlighting how each LLM's accuracy influenced overall predictions. The analysis used Python packages, including pandas 2.1.4, numpy 1.24.3, scikit-learn 1.3.0, scipy 1.11.4, catboost 1.2, LightGBM 4.1.0, seaborn 0.12.2, SHAP 0.42.1, and matplotlib 3.8.0.

Results

Distribution of Question Types in the CNNLE Over the Years

Figure 1 illustrates the distribution of question types over the years in both sections of the CNNLE. Figure 1A depicts the distribution of question types over the years in the Practical Skills section. In the Practical Skills section, A1-type questions decreased from 86 in 2019 to 59 in 2023, while A2-type questions increased from 18 to 43. A3-type and A4-type questions showed smaller fluctuations. Figure 1B shows the distribution of question types over the years in the Professional Practice section. In the Professional Practice section, A1-type questions fell from 67 in 2019 to 55 in 2023, while A2-type questions increased from 33 to 45. A3-type questions remained relatively stable, with minor variations.

Figure 1. Distribution of question types in CNNLE Professional Practice and Practical Skills sections (2019-2023). (A) Distribution of question types in Practical Skills from 2019 to 2023. (B) Distribution of question types in Professional Practice from 2019 to 2023.



Accuracy of LLMs in Professional Practice

Table 1 presents the accuracy of LLMs in the Professional Practice section from 2019 to 2023. In 2023, Qwen-2.5 achieved the highest accuracy (0.850), followed by ERNIE Bot-3.5 (0.808) and GPT-4o (0.783). GPT-4.0 consistently outperformed

GPT-3.5 in all years, with scores of 0.725 and 0.492, respectively, in 2023. Copilot and SPARK also showed moderate performance improvements over time, reaching 0.775 and 0.692 in 2023. Across the 5 years, Qwen-2.5 demonstrated the best overall accuracy (0.875), followed by GPT-4o (0.803) and ERNIE Bot-3.5 (0.785).

Table 1. Accuracy of large language models in Professional Practice.

Year	GPT-3.5	GPT-4.0	GPT-4o	Copilot	ERNIE Bot-3.5	SPARK	Qwen-2.5
2023	0.492	0.725	0.783	0.775	0.808	0.692	0.850
2022	0.450	0.675	0.833	0.767	0.758	0.667	0.917
2021	0.517	0.683	0.817	0.725	0.783	0.650	0.900
2020	0.500	0.708	0.725	0.733	0.767	0.600	0.850
2019	0.550	0.758	0.858	0.583	0.808	0.600	0.858
Overall	0.502	0.710	0.803	0.717	0.785	0.642	0.875

Accuracy of LLMs in Practical Skills

Table 2 presents the accuracy of LLMs in the Practical Skills section from 2019 to 2023. In 2023, Qwen-2.5 achieved the highest accuracy (0.908), followed by GPT-4o (0.833) and Copilot (0.792). GPT-4.0 and ERNIE Bot-3.5 both scored 0.775,

showing steady improvement compared with earlier years. SPARK and GPT-3.5 performed moderately, with scores of 0.758 and 0.550, respectively. Over the 5 years, Qwen-2.5 consistently outperformed other models, achieving the highest overall accuracy (0.903). GPT-4o followed with 0.810, while ERNIE Bot-3.5 ranked third with 0.777.

Table 2. Accuracy of large language models in Practical Skills.

Year	GPT-3.5	GPT-4.0	GPT-4o	Copilot	ERNIE Bot-3.5	SPARK	Qwen-2.5
2023	0.550	0.775	0.833	0.792	0.775	0.758	0.908
2022	0.467	0.692	0.800	0.792	0.792	0.675	0.850
2021	0.467	0.708	0.850	0.667	0.750	0.567	0.942
2020	0.475	0.642	0.783	0.592	0.800	0.700	0.933
2019	0.483	0.658	0.783	0.458	0.767	0.592	0.883
Overall	0.488	0.695	0.810	0.660	0.777	0.658	0.903

Accuracy of LLMs for Question Types

Table 3 indicates the accuracy of LLMs across 4 question types (A1, A2, A3, and A4) from 2019 to 2023. In 2023, Qwen-2.5 achieved the highest accuracy for A1, A2, and A3 questions (0.860, 0.909, and 0.853, respectively), while all models reached perfect accuracy (1.000) for A4 questions. GPT-4o consistently performed well across all question types, ranking second or

third in accuracy. In 2022, Qwen-2.5 maintained high performance across A1, A2, and A3 questions (0.913, 0.810, and 0.963, respectively). From 2019 to 2021, Qwen-2.5 demonstrated steady improvements across all question types. Overall, Qwen-2.5 achieved the highest average accuracy (0.889), followed by GPT-4o (0.807) and ERNIE Bot-3.5 (0.781).

Table 3. Accuracy of large language models for 4 question types.

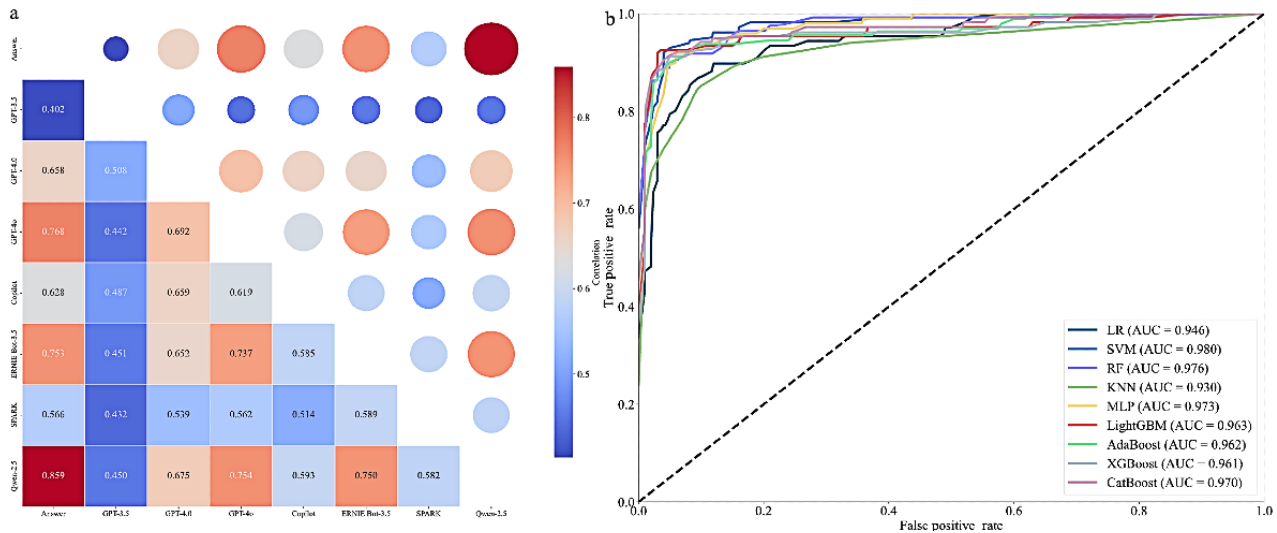
Question type	GPT-3.5	GPT-4.0	GPT-4o	Copilot	ERNIE Bot-3.5	SPARK	Qwen-2.5
2023							
A1	0.526	0.684	0.789	0.763	0.763	0.719	0.860
A2	0.443	0.807	0.830	0.784	0.807	0.705	0.909
A3	0.647	0.794	0.794	0.824	0.824	0.765	0.853
A4	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2022							
A1	0.476	0.746	0.881	0.825	0.817	0.698	0.913
A2	0.405	0.582	0.671	0.696	0.671	0.620	0.810
A3	0.519	0.667	0.889	0.778	0.852	0.630	0.963
A4	0.500	0.750	1.000	0.875	0.875	0.875	0.875
2021							
A1	0.467	0.660	0.820	0.667	0.727	0.600	0.927
A2	0.538	0.738	0.846	0.738	0.831	0.646	0.938
A3	0.524	0.762	0.857	0.714	0.857	0.571	0.810
A4	0.500	1.000	1.000	1.000	0.750	0.500	1.000
2020							
A1	0.478	0.675	0.771	0.650	0.771	0.656	0.879
A2	0.492	0.695	0.763	0.695	0.814	0.678	0.915
A3	0.550	0.600	0.550	0.650	0.750	0.600	0.900
A4	0.500	0.750	1.000	0.750	1.000	0.250	1.000
2019							
A1	0.490	0.680	0.804	0.503	0.791	0.601	0.869
A2	0.569	0.745	0.843	0.588	0.784	0.588	0.843
A3	0.556	0.778	0.861	0.500	0.778	0.583	0.917
Overall	0.495	0.703	0.807	0.688	0.781	0.650	0.889

Correlation Heatmap and AUC Curves Using Machine Learning

Figure 2 provides an analysis of the correlation heatmap and AUC curves for machine learning models. Figure 2A presents the correlation heatmap, where Qwen-2.5 shows the highest correlation with correct answers ($r=0.859$), while GPT-3.5

shows the lowest correlation ($r=0.402$). Figure 2B illustrates the AUC scores for each machine learning model in the multiclass classification task. The models achieved the following AUC scores: LR (0.946), SVM (0.980), RF (0.976), KNN (0.930), MLP (0.973), LightGBM (0.963), AdaBoost (0.962), XGBoost (0.961), and CatBoost (0.970).

Figure 2. Correlation heatmap and AUC curves of machine learning models in CNNLE. (A) Correlation heatmap: The heatmap illustrates the relationships between different LLMs. The lower left displays numerical correlation values, while the upper right represents correlation magnitude through circle size. Color gradients range from blue (low correlation) to red (high correlation), providing a visual summary of metric interdependencies. (B) AUC curves: The AUC curves compare the performance of various machine learning and ensemble models, highlighting their classification accuracy across the data set. AdaBoost: Adaptive Boosting; AUC: area under the curve; CatBoost: Categorical Boosting; KNN: k-nearest neighbor; LightGBM: Light Gradient-Boosting Machine; LR: Logistic Regression; MLP: Multilayer Perceptron; RF: Random Forest; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boosting.



Metrics for 5-Class Classification Using Machine Learning

Table 4 presents a comparative analysis of 9 machine learning models for multiclass classification, evaluated by average metrics including AUC, accuracy, sensitivity, specificity,

precision, PPV, F_1 -score, and NPV. Among these, the SVM and XGBoost models achieve AUC values of 0.980 and 0.961, along with accuracy scores of 0.858 and 0.908, respectively. In contrast, LR and KNN exhibit lower accuracy scores of 0.817 and 0.767.

Table 4. Metrics of machine learning.

Classifier	AUC ^a	Accuracy	Sensitivity	Specificity	Precision	PPV ^b	F ₁ -score	NPV ^c
LR ^d	0.946	0.817	0.808	0.953	0.818	0.818	0.809	0.954
SVM ^e	0.980	0.858	0.857	0.965	0.861	0.861	0.854	0.964
RF ^f	0.976	0.858	0.860	0.965	0.856	0.856	0.854	0.964
KNN ^g	0.930	0.767	0.772	0.942	0.787	0.787	0.768	0.941
MLP ^h	0.973	0.825	0.823	0.957	0.830	0.830	0.819	0.956
LightGBM ⁱ	0.963	0.900	0.908	0.975	0.895	0.895	0.899	0.974
AdaBoost ^j	0.962	0.858	0.859	0.964	0.855	0.855	0.856	0.964
XGBoost ^k	0.961	0.908	0.905	0.978	0.901	0.901	0.901	0.977
CatBoost ^l	0.970	0.892	0.892	0.974	0.885	0.885	0.885	0.973

^aAUC: area under the curve.

^bPPV: positive predictive value.

^cNPV: negative predictive value.

^dLR: Logistic Regression.

^eSVM: Support Vector Machine.

^fRF: Random Forest.

^gKNN: k-nearest neighbor.

^hMLP: Multilayer Perceptron.

ⁱLightGBM: Light Gradient-Boosting Machine.

^jAdaBoost: Adaptive Boosting.

^kXGBoost: Extreme Gradient Boosting.

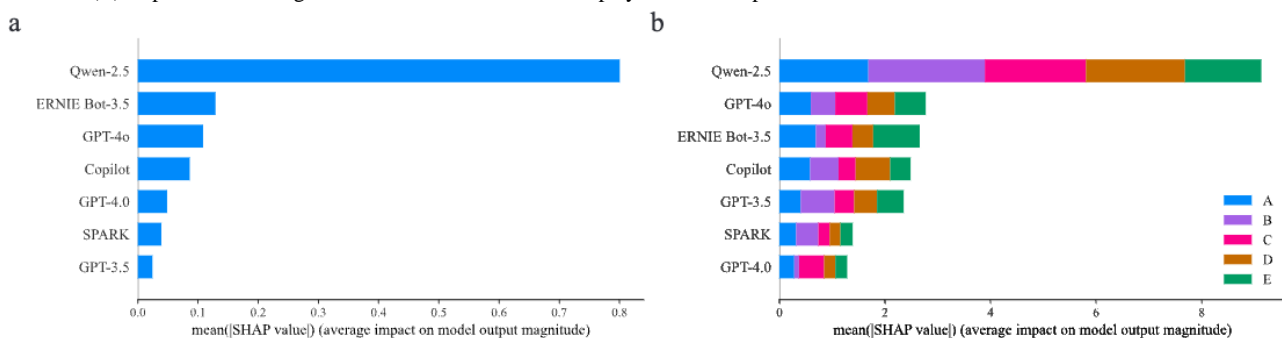
^lCatBoost: Categorical Boosting.

Importance Ranking of SVM and XGBoost Models

Figure 3 presents the SHAP summary bar plot for both the SVM and XGBoost models. In Figure 3A, the SVM model ranks the features as follows: Qwen-2.5, ERNIE Bot-3.5, GPT-4o, Copilot, GPT-4.0, SPARK, and GPT-3.5. Meanwhile, Figure

3B shows the importance ranking of XGBoost model with a slightly different order: Qwen-2.5, GPT-4o, ERNIE Bot-3.5, Copilot, GPT-3.5, SPARK, and GPT-4.0. Qwen-2.5 stands out as the most influential feature in both models. Furthermore, including other LLMs enhances overall model performance, evidenced by improvements in AUC and accuracy.

Figure 3. SHAP summary bar plot in Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost) models. (A) Importance ranking of SVM model. (B) Importance ranking of XGBoost model. SHAP: Shapley Additive Explanations.



Discussion

Principal Findings

This study is the first to evaluate the performance of 7 LLMs on the CNNLE dataset (2019-2023), highlighting significant advancements in Chinese LLM development and their applications in nursing education. Among the models tested,

Qwen-2.5 demonstrated the highest accuracy (88.92%), significantly surpassing the performance of the other LLMs. This superior accuracy can be attributed to its training on an extensive Chinese dataset and optimized parameters, enabling it to handle domain-specific nursing knowledge and complex clinical decision-making tasks with exceptional precision. These results underline the growing feasibility of deploying advanced LLMs such as Qwen-2.5 to support standardized nursing

examinations in China, offering consistent, scalable, and efficient assessments.

Our findings offer a clear pathway for the practical application of Qwen-2.5 in nursing curricula and professional training. For instance, Qwen-2.5 could serve as a virtual tutor, providing personalized feedback and explanations to nursing students in real time. Its ability to respond promptly and accurately to a wide range of questions makes it particularly valuable for addressing individual knowledge gaps and reinforcing complex concepts. Educators could incorporate Qwen-2.5 into classroom activities, using it to simulate clinical scenarios or evaluate students' decision-making skills. Furthermore, mobile apps powered by Qwen-2.5 could allow nursing students to access high-quality, interactive learning resources anytime and anywhere, thereby enhancing accessibility and flexibility in education. Beyond supporting student learning, Qwen-2.5 and other LLMs can enhance professional development for practicing nurses. For instance, these models could be integrated into continuing education programs, where they act as interactive resources to update practitioners on the latest evidence-based practices. By serving as a knowledge repository, LLMs can enable nurses to quickly access relevant guidelines, ensuring timely and informed clinical decisions.

The results also extend prior research by demonstrating how ensemble machine learning methods can enhance LLM performance in specialized tasks. By integrating the outputs of 7 LLMs using the XGBoost algorithm, we achieved an improved accuracy of 90.83%, surpassing the best-performing single model. This novel application of ensemble methods highlights a promising direction for developing personalized LLMs tailored to specific domains, such as health care education. Previous studies, including those by Li et al [38] and Brin et al [39], have emphasized the value of context-specific tuning, but our research provides concrete evidence of the effectiveness of combining multiple models to enhance accuracy in domain-specific applications.

Furthermore, our study situates Chinese LLMs within the broader global landscape of AI. While skepticism has persisted regarding whether Chinese LLMs can rival models developed by OpenAI or Google, our results demonstrate that Qwen-2.5 not only excels on the CNNLE assessment but also outperforms other LLMs. For example, Qwen-2.5 achieved a higher accuracy than GPT-4 (72.5%) on similar standardized tests, as reported in previous studies [40]. This performance underscores the competitive edge of Chinese-developed models, particularly in addressing language-specific and cultural nuances in health care education.

Our findings also reveal recent trends in nursing education, such as the increasing complexity of standardized examinations such as the CNNLE, which has transitioned from straightforward A1-type questions to more analytical A2-type clinical case scenarios. This shift reflects the growing need for nursing professionals to develop higher-order reasoning skills and apply clinical knowledge to real-world situations. Qwen-2.5's ability to process nuanced clinical scenarios positions it as an effective tool for addressing these demands. By integrating models such as Qwen-2.5 into nursing curricula, educators can better prepare

students for the complexities of modern health care through scenario-based learning and real-time feedback.

Our study also demonstrates the broader impact of China's advancements in AI, particularly through open access LLMs such as Qwen-2.5 and ERNIE Bot-3.5, which provide practical solutions for addressing regional disparities in nursing education. These models are especially valuable in regions where access to global LLMs, such as OpenAI's GPT models, is restricted, as they deliver high-quality, localized educational content. By promoting the standardization of nursing education across institutions, these tools help bridge resource gaps and improve the overall quality of health care training in China. Furthermore, their ability to enable personalized and flexible learning through mobile apps empowers students to access educational resources anytime and anywhere. This adaptability positions Chinese LLMs as critical tools for advancing specialized education while addressing both regional inequalities and broader challenges in health care training.

Limitations

First, our evaluation relied on MCQs to assess the knowledge of LLMs, which may not fully capture their ability to handle open-ended or complex clinical tasks. Future studies could incorporate open-ended questions, clinical simulations, or case-based assessments to evaluate LLMs' reasoning and decision-making capabilities more comprehensively. These methods would better reflect the unstructured and nuanced scenarios encountered in real-world clinical practice, providing a deeper understanding of how LLMs process complex clinical information. Second, the performance of LLMs can vary based on factors such as prompt design, the number of questions asked, and the context of those questions, introducing variability into results. To address this, standardized evaluation protocols should be developed to ensure consistency across benchmarking studies. Furthermore, future research could focus on refining prompt engineering techniques and optimizing model fine-tuning to improve accuracy and reliability in diverse clinical applications. These refinements could support the development of LLMs that are better suited to handling complex scenarios, such as differential diagnoses or multistep decision-making. Third, while Qwen-2.5 demonstrated highest accuracy on the CNNLE dataset, its optimization for the Chinese language and MCQ format may limit its generalizability to other medical domains and contexts. Future studies should evaluate its applicability in multilingual and open-ended settings to assess its effectiveness in tasks beyond standardized testing formats and within various health care contexts. To enhance the suitability of LLMs for specialized health care tasks, such as diagnostic reasoning and treatment planning, future research could prioritize the development of domain-specific models. This could involve fine-tuning LLMs on datasets that include detailed case histories, diagnostic pathways, and clinical protocols. Such datasets would allow the models to learn context-specific patterns and reasoning processes, equipping them to provide more accurate and relevant recommendations in clinical settings. Furthermore, fine-tuned models could be used to assist in treatment planning by integrating data from clinical guidelines, patient histories, and risk assessment tools to offer tailored suggestions for patient care. Addressing biases in LLM training is also essential for

ensuring equitable decision-making across diverse patient populations. Researchers should consider incorporating fairness-aware algorithms and curated datasets that reflect demographic diversity to mitigate potential biases. Such efforts could ensure that domain-specific LLMs provide consistent and unbiased recommendations, particularly in high-stakes environments such as emergency department triage workflows. Finally, this study focused on general purpose LLMs, excluding models explicitly trained for medical tasks, such as Gemini or Claude. Preliminary findings suggest that fine-tuned medical models may achieve superior accuracy for specific applications. Future research should conduct comparative evaluations of general purpose and domain-specific LLMs to identify the

optimal approach for different health care needs. These studies could also assess whether fine-tuned models are more effective in real-time clinical workflows, such as triage systems, or in supporting complex decision-making across various medical specialties.

Conclusions

This study is the first to evaluate the performance of 7 LLMs on the CNNLE and that the integration of models via machine learning significantly boosted accuracy, reaching 90.8%. These findings demonstrate the transformative potential of LLMs in revolutionizing health care education and call for further research to refine their capabilities and expand their impact on examination preparation and professional training.

Authors' Contributions

SZ was responsible for the conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing of the original draft, reviewing, editing the manuscript, and visualization. WH and ZY managed data curation and visualization. JY also handled data curation and contributed to reviewing and editing the manuscript. FZ supervised the project, contributed to reviewing and editing the manuscript, and managed project administration. All authors have read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work, ensuring that any questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Original data set.

[[XLSX File \(Microsoft Excel File\), 68 KB-Multimedia Appendix 1](#)]

Multimedia Appendix 2

Ipybn notebook codes.

[[PDF File \(Adobe PDF File\), 249 KB-Multimedia Appendix 2](#)]

References

1. American Association of Critical-Care Nurses. Safeguarding the patient and the profession: the value of critical care nurse certification. *Am J Crit Care*. 2003;12(2):154-164. [Medline: [12625174](#)]
2. Gambrell E. *Critical Thinking in Clinical Practice: Improving the Quality of Judgments and Decisions*. Hoboken, NJ. John Wiley & Sons; 2006.
3. 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 Med Educ*. 2024;24(1):143. [[FREE Full text](#)] [doi: [10.1186/s12909-024-05125-7](#)] [Medline: [38355517](#)]
4. Miao Y, Luo Y, Zhao Y, Li J, Liu M, Wang H, et al. Performance of GPT-4 on Chinese nursing examination: potentials for AI-assisted nursing education using large Language models. *Nurse Educ*. 2024;49(6):E338-E343. [doi: [10.1097/NNE.0000000000001679](#)] [Medline: [38981035](#)]
5. Allen C, Woodnutt S. Can ChatGPT pass a nursing exam? *Int J Nurs Stud*. 2023;145:104522. [doi: [10.1016/j.ijnurstu.2023.104522](#)] [Medline: [37354792](#)]
6. Castonguay A, Farthing P, Davies S, Vogelsang L, Kleib M, Risling T, et al. Revolutionizing nursing education through ai integration: a reflection on the disruptive impact of ChatGPT. *Nurse Educ Today*. 2023;129:105916. [doi: [10.1016/j.nedt.2023.105916](#)] [Medline: [37515957](#)]
7. Chang BS. Transformation of undergraduate medical education in 2023. *JAMA*. 2023;330(16):1521-1522. [doi: [10.1001/jama.2023.16943](#)] [Medline: [37698855](#)]
8. Branum C, Schiavenato M. Can ChatGPT accurately answer a PICOT question? Assessing ai response to a clinical question. *Nurse Educ*. 2023;48(5):231-233. [doi: [10.1097/NNE.0000000000001436](#)] [Medline: [37130197](#)]
9. Harris E. Large language models answer medical questions accurately, but can't match clinicians' knowledge. *JAMA*. 2023;330(9):792-794. [doi: [10.1001/jama.2023.14311](#)] [Medline: [37548971](#)]

10. Alam F, Lim MA, Zulkipli IN. Integrating AI in medical education: embracing ethical usage and critical understanding. *Front Med (Lausanne)*. 2023;10:1279707. [FREE Full text] [doi: [10.3389/fmed.2023.1279707](https://doi.org/10.3389/fmed.2023.1279707)] [Medline: [37901398](https://pubmed.ncbi.nlm.nih.gov/37901398/)]
11. Yang H. How I use ChatGPT responsibly in my teaching. *Nature*. 2023;(476-4687):1-3. [FREE Full text] [doi: [10.1038/d41586-023-01026-9](https://doi.org/10.1038/d41586-023-01026-9)] [Medline: [37045954](https://pubmed.ncbi.nlm.nih.gov/37045954/)]
12. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. *ACM Trans Intell Syst Technol*. 2024;15(3):1-45. [doi: [10.1145/3641289](https://doi.org/10.1145/3641289)]
13. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digit Med*. 2024;7(1):82. [FREE Full text] [doi: [10.1038/s41746-024-01074-z](https://doi.org/10.1038/s41746-024-01074-z)] [Medline: [38553625](https://pubmed.ncbi.nlm.nih.gov/38553625/)]
14. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. [FREE Full text] [doi: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)] [Medline: [37438534](https://pubmed.ncbi.nlm.nih.gov/37438534/)]
15. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. [doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)] [Medline: [37460753](https://pubmed.ncbi.nlm.nih.gov/37460753/)]
16. Moura L, Jones DT, Sheikh IS, Murphy S, Kalfin M, Kummer BR, et al. Implications of large language models for quality and efficiency of neurologic care: emerging issues in neurology. *Neurology*. 2024;102(11):e209497. [doi: [10.1212/WNL.0000000000209497](https://doi.org/10.1212/WNL.0000000000209497)] [Medline: [38759131](https://pubmed.ncbi.nlm.nih.gov/38759131/)]
17. Kernberg A, Gold JA, Mohan V. Using ChatGPT-4 to create structured medical notes from audio recordings of physician-patient encounters: comparative study. *J Med Internet Res*. 2024;26:e54419. [FREE Full text] [doi: [10.2196/54419](https://doi.org/10.2196/54419)] [Medline: [38648636](https://pubmed.ncbi.nlm.nih.gov/38648636/)]
18. Caruccio L, Cirillo S, Polese G, Solimando G, Sundaramurthy S, Tortora G. Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. *Expert Syst Appl*. 2024;235:121186. [doi: [10.1016/j.eswa.2023.121186](https://doi.org/10.1016/j.eswa.2023.121186)]
19. Ferdush J, Begum M, Hossain ST. ChatGPT and clinical decision support: scope, application, and limitations. *Ann Biomed Eng*. 2024;52(5):1119-1124. [doi: [10.1007/s10439-023-03329-4](https://doi.org/10.1007/s10439-023-03329-4)] [Medline: [37516680](https://pubmed.ncbi.nlm.nih.gov/37516680/)]
20. Su MC, Lin LE, Lin LH, Chen YC. Assessing question characteristic influences on ChatGPT's performance and response-explanation consistency: insights from taiwan's nursing licensing exam. *Int J Nurs Stud*. 2024;153:104717. [doi: [10.1016/j.ijnurstu.2024.104717](https://doi.org/10.1016/j.ijnurstu.2024.104717)] [Medline: [38401366](https://pubmed.ncbi.nlm.nih.gov/38401366/)]
21. Pham C, Govender R, Tehami S, Chavez S, Adepoju OE, Liaw W. ChatGPT's performance in cardiac arrest and bradycardia simulations using the American Heart Association's advanced cardiovascular life support guidelines: exploratory study. *J Med Internet Res*. 2024;26:e55037. [FREE Full text] [doi: [10.2196/55037](https://doi.org/10.2196/55037)] [Medline: [38648098](https://pubmed.ncbi.nlm.nih.gov/38648098/)]
22. Burk-Rafel J, Reinstein I, Feng J, Kim MB, Miller LH, Cocks PM, et al. Development and validation of a machine learning-based decision support tool for residency applicant screening and review. *Acad Med*. 2021;96(11S):S54-S61. [doi: [10.1097/ACM.0000000000004317](https://doi.org/10.1097/ACM.0000000000004317)] [Medline: [34348383](https://pubmed.ncbi.nlm.nih.gov/34348383/)]
23. Wang D, Liang J, Ye J, Li J, Li J, Zhang Q, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. *J Med Internet Res*. 2024;26:e58041. [FREE Full text] [doi: [10.2196/58041](https://doi.org/10.2196/58041)] [Medline: [39046096](https://pubmed.ncbi.nlm.nih.gov/39046096/)]
24. Li S. Exploring the clinical capabilities and limitations of ChatGPT: a cautionary tale for medical applications. *Int J Surg*. 2023;109(9):2865-2867. [FREE Full text] [doi: [10.1097/JS9.0000000000000523](https://doi.org/10.1097/JS9.0000000000000523)] [Medline: [37222684](https://pubmed.ncbi.nlm.nih.gov/37222684/)]
25. Jiang Y, Hao J, Fauss M, Li C. Detecting ChatGPT-generated essays in a large-scale writing assessment: is there a bias against non-native english speakers? *Comput Educ*. 2024;217:105070. [doi: [10.1016/j.compedu.2024.105070](https://doi.org/10.1016/j.compedu.2024.105070)]
26. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health*. 2024;6(1):e12-e22. [FREE Full text] [doi: [10.1016/S2589-7500\(23\)00225-X](https://doi.org/10.1016/S2589-7500(23)00225-X)] [Medline: [38123252](https://pubmed.ncbi.nlm.nih.gov/38123252/)]
27. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nat Commun*. 2024;15(1):2050. [FREE Full text] [doi: [10.1038/s41467-024-46411-8](https://doi.org/10.1038/s41467-024-46411-8)] [Medline: [38448475](https://pubmed.ncbi.nlm.nih.gov/38448475/)]
28. Hou J, Chen S, Sabharwal S, Fan V, Yan M, Wang W. Comparison of RN licensure examination: China and the United States. *Int J Nurs Sci*. 2019;6(1):111-116. [FREE Full text] [doi: [10.1016/j.ijnss.2018.11.002](https://doi.org/10.1016/j.ijnss.2018.11.002)] [Medline: [31406876](https://pubmed.ncbi.nlm.nih.gov/31406876/)]
29. Nusinovi S, Tham YC, Chak Yan MY, Wei Ting DS, Li J, Sabanayagam C, et al. Logistic regression was as good as machine learning for predicting major chronic diseases. *J Clin Epidemiol*. 2020;122:56-69. [doi: [10.1016/j.jclinepi.2020.03.002](https://doi.org/10.1016/j.jclinepi.2020.03.002)]
30. Suthaharan S. Support vector machine. In: *Machine Learning Models and Algorithms for Big Data Classification: Thinking With Examples for Effective Learning*. New York, NY: Springer; 2016:207-235.
31. Tang J, Deng C, Huang GB. Extreme learning machine for multilayer perceptron. *IEEE Trans Neural Netw Learn Syst*. 2016;27(4):809-821. [doi: [10.1109/TNNLS.2015.2424995](https://doi.org/10.1109/TNNLS.2015.2424995)] [Medline: [25966483](https://pubmed.ncbi.nlm.nih.gov/25966483/)]
32. Samet H. K-nearest neighbor finding using MaxNearestDist. *IEEE Trans Pattern Anal Mach Intell*. 2008;30(2):243-252. [doi: [10.1109/TPAMI.2007.1182](https://doi.org/10.1109/TPAMI.2007.1182)] [Medline: [18084056](https://pubmed.ncbi.nlm.nih.gov/18084056/)]
33. Segal MR. Machine learning benchmarks and random forest regression. 2004. URL: <https://escholarship.org/uc/item/35x3v9t4> [accessed 2004-04-14]

34. Shehadeh A, Alshboul O, Al Mamlook RE, Hamedat O. Machine learning models for predicting the residual value of heavy construction equipment: an evaluation of modified decision tree, LightGBM, and XGBoost regression. *Automation Constr.* 2021;129:103827. [doi: [10.1016/j.autcon.2021.103827](https://doi.org/10.1016/j.autcon.2021.103827)]
35. Zhang PB, Yang ZX. A novel adaBoost framework with robust threshold and structural optimization. *IEEE Trans Cybern.* 2018;48(1):64-76. [doi: [10.1109/TCYB.2016.2623900](https://doi.org/10.1109/TCYB.2016.2623900)] [Medline: [27898387](https://pubmed.ncbi.nlm.nih.gov/27898387/)]
36. Chen T, Guestrin C. A scalable tree boosting system. 2016. Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and data Mining; 2016 August 17; San Francisco, CA. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
37. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. *J Big Data.* 2020;7(1):94. [FREE Full text] [doi: [10.1186/s40537-020-00369-8](https://doi.org/10.1186/s40537-020-00369-8)] [Medline: [33169094](https://pubmed.ncbi.nlm.nih.gov/33169094/)]
38. Li DJ, Kao YC, Tsai SJ, Bai YM, Yeh TC, Chu CS, et al. Comparing the performance of ChatGPT GPT-4, Bard, and llama-2 in the Taiwan Psychiatric Licensing Examination and in differential diagnosis with multi-center psychiatrists. *Psychiatry Clin Neurosci.* 2024;78(6):347-352. [doi: [10.1111/pcn.13656](https://doi.org/10.1111/pcn.13656)] [Medline: [38404249](https://pubmed.ncbi.nlm.nih.gov/38404249/)]
39. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep.* 2023;13(1):16492. [FREE Full text] [doi: [10.1038/s41598-023-43436-9](https://doi.org/10.1038/s41598-023-43436-9)] [Medline: [37779171](https://pubmed.ncbi.nlm.nih.gov/37779171/)]
40. Sato H, Ogasawara K. ChatGPT (GPT-4) passed the Japanese national license examination for pharmacists in 2022, answering all items including those with diagrams: a descriptive study. *J Educ Eval Health Prof.* 2024;21:4. [FREE Full text] [doi: [10.3352/jeehp.2024.21.4](https://doi.org/10.3352/jeehp.2024.21.4)] [Medline: [38413129](https://pubmed.ncbi.nlm.nih.gov/38413129/)]

Abbreviations

AdaBoost: Adaptive Boosting
AI: artificial intelligence
AUC: area under the curve
CNNLE: Chinese National Nursing Licensing Examination
KNN: k-nearest neighbors
LightGBM: Light Gradient-Boosting Machine
LLM: large language model
LR: Logistic Regression
MCQ: multiple-choice question
MLP: Multilayer Perceptron
NPV: negative predictive value
PPV: positive predictive value
RF: Random Forest
SHAP: Shapley Additive Explanations
SVM: Support Vector Machine
XGBoost: Extreme Gradient Boosting

Edited by A Castonguay; submitted 27.06.24; peer-reviewed by Y Hirano, M Besler; comments to author 26.07.24; revised version received 06.08.24; accepted 20.12.24; published 10.01.25

Please cite as:

Zhu S, Hu W, Yang Z, Yan J, Zhang F

Qwen-2.5 Outperforms Other Large Language Models in the Chinese National Nursing Licensing Examination: Retrospective Cross-Sectional Comparative Study

JMIR Med Inform 2025;13:e63731

URL: <https://medinform.jmir.org/2025/1/e63731>

doi: [10.2196/63731](https://doi.org/10.2196/63731)

PMID:

©Shiben Zhu, Wanqin Hu, Zhi Yang, Jiani Yan, Fang Zhang. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org/>), 10.01.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.