Preplanned Studies

Evaluating Large Language Models' Potential in Field Epidemiology Investigation Based on Chinese Context — Zhejiang Province, China, 2025

Tao Zhang[1,&]; Qifeng Zhao[2,&]; Yaxin Dai[3]; Mengna Wu[1]; Yujia Zhai[1]; Le Xu[1]; Xue Gu[1]; Junfen Lin[1]; Chen Wu[1,#]

Summary

What is already known about this topic?

Large language models (LLMs) have demonstrated considerable potential in clinical applications. However, their performance in field epidemiology, particularly within Chinese-language contexts, remains largely unexplored.

What is added by this report?

This study evaluates six leading LLMs (ChatGPT-o4-mini-high, ChatGPT-4o, DeepSeek-R1, DeepSeek-V3, Qwen3-235B-A22B, and Qwen2.5-Max) using examination questions from the Zhejiang Field Epidemiology Training Program. For multiple-choice questions, all models except DeepSeek-V3 scored below the 75th percentile of junior field epidemiologists, while for case-based questions, LLMs generally outperformed that percentile. However, LLMs demonstrated significant limitations when addressing questions requiring specialized knowledge. Notably, LLMs may generate inaccurate or fabricated references, presenting substantial risks for inexperienced practitioners.

What are the implications for public health practice?

LLMs demonstrate promising potential for supporting epidemiological investigations. Nevertheless, current LLMs cannot replace human expertise in field epidemiology. Their practical implementation faces considerable challenges, including ensuring output accuracy and reliability. Future efforts should prioritize optimizing performance through verified knowledge databases and establishing robust regulatory frameworks to enhance their effectiveness in public health applications.

ABSTRACT

Introduction Large language models (LLMs) have demonstrated potential applications across diverse fields, yet their effectiveness in supporting field epidemiology investigations remains uncertain.

Methods We assessed six prominent LLMs (ChatGPT-o4-mini-high, ChatGPT-4o, DeepSeek-R1, DeepSeek-V3, Qwen3-235B-A22B, and Qwen2.5-max) using multiple-choice and case-based questions from the 2025 Zhejiang Field Epidemiology Training Program entrance examination. Model responses were evaluated against standard answers and benchmarked against performance scores from junior epidemiologists.

Results For multiple-choice questions, only DeepSeek-V3 (75%) exceeded the 75th percentile performance level of junior epidemiologists (67.5%). In case-based assessments, most LLMs achieved or surpassed the 75th percentile of junior epidemiologists, demonstrating particular strength in data analysis tasks.

Conclusion Although LLMs demonstrate promise as supportive tools in field epidemiology investigations, they cannot yet replace human expertise. Significant challenges persist regarding the accuracy and timeliness of model outputs, alongside critical concerns about data security and privacy protection that must be addressed before widespread implementation.

Field epidemiology investigation serves as a cornerstone of public health practice, proving essential for identifying risk factors and implementing effective control measures. Large language models (LLMs) have recently emerged as potentially transformative tools in this domain ([1]). Models such as ChatGPT and DeepSeek have demonstrated impressive capabilities in text generation, reasoning, and data analysis. These systems can interpret user commands and generate contextually appropriate responses, positioning LLMs as valuable support tools across diverse fields.

Previous research has primarily concentrated on clinical applications of LLMs, where they have shown promise in medical diagnosis, patient counseling, and medical record management (2). While these applications highlight the broad potential of LLMs, their effectiveness in supporting field epidemiology investigations remains uncertain. Field epidemiology investigation encompasses extensive knowledge domains, including clinical medicine, epidemiology, laboratory and behavioral sciences, laws and regulations, technical guidelines, and decision-making frameworks (3). The existing literature on LLMs in public health remains limited, with few studies specifically examining their role in field epidemiology investigations. Moreover, most research has been conducted in Western contexts, leaving the application of LLMs in field epidemiology investigations — particularly within Chinese-language environments — largely unexplored. Given the rapid advancement of artificial intelligence (AI) Plus initiatives, investigating how LLMs can assist epidemiological investigations carries significant practical importance.

This study addressed this knowledge gap by evaluating the performance of several leading LLMs in executing common field epidemiology investigation tasks. The research not only contributes to a broader understanding of LLM applications in public health but also provides valuable insights for developing AI-assisted tools for field epidemiology investigations in China.

We selected six leading large language models for evaluation: three reasoning models (ChatGPT-o4-mini-high, DeepSeek-R1, and Qwen3-235B-A22B) and three non-reasoning models (DeepSeek-V3, Qwen2.5-max, and ChatGPT-4o). ChatGPT-o4-mini-high and ChatGPT-4o are proprietary closed-source models, while the remaining four represent open-source alternatives. Our evaluation framework utilized questions from the 2025 Zhejiang Field Epidemiology Training Program entrance examination, with all materials reviewed by field epidemiology experts to ensure accuracy and clarity. A total of 35 junior field epidemiologists participated in the examination. The assessment comprised two components: multiple-choice questions testing foundational knowledge and case-based scenarios evaluating practical application skills. The multiple-choice section included 20 single-answer questions with five options each, covering core topics such as infectious disease surveillance and reporting, risk assessment, outbreak management protocols, and sample collection procedures. The case-based questions presented open-ended scenarios requiring sequential responses, with each subsequent question posed only after the model completed its previous answer. This approach simulates real-world outbreak response conditions and evaluates the models' capacity to provide accurate, professional guidance on demand. All models were accessed on May 12, 2025, through their respective web interfaces using standardized Chinese-language prompts. Additional methodological details are available in the Supplementary Material (available at https://weekly.chinacdc.cn/).

For multiple-choice questions, we compared model responses against standard answers, awarding one point for each correct response (maximum score: 20 points). The case-based section contained four questions, with each response independently evaluated by two expert assessors. These evaluators scored responses against established criteria, including scientific accuracy, comprehensiveness, clarity of presentation, and contextual relevance. Each open-ended question carried a maximum score of 10 points.

For the multiple-choice questions, we calculated the proportion of correct answers for each LLM and compared these results with responses from junior epidemiologists. Statistical differences were assessed using binomial tests with p_0=0.20 (LLMs versus chance) and bootstrap approaches (highest-scoring LLM versus junior epidemiologists). For the case-based questions, we computed Pearson's *r* and Spearman's ρ to evaluate the correlation between the two evaluators' ratings. We conducted Friedman and Wilcoxon tests to examine score differences in the open-ended questions. All statistical analyses were performed using the "stats" package in R software (version 4.3.2, R Core Team, Vienna, Austria). Statistical significance was set at P≤0.05.

Figure 1 demonstrates the performance of each LLM on the multiple-choice questions. Among the 20 questions, DeepSeek-V3 and Qwen3-235B-A22B achieved the highest scores, with 15/20 [75%, 95% confidence interval (*CI*): 50.9%, 91.3%] and 13/20 (65%, 95% *CI*: 40.8%, 84.6%), respectively. ChatGPT-o4-mini-high and ChatGPT-4o obtained the lowest scores, both scoring 8/20 (40%, 95% *CI*: 19.1%, 63.9%). The results revealed that four models achieved accuracy rates significantly higher than random guessing (P<0.05), with the exceptions being ChatGPT-o4-mini-high and ChatGPT-4o. Additional results are provided in the supplementary materials (Figure S1 and Table S1). When comparing the top-performing model, DeepSeek-V3 demonstrated significantly better performance than the median

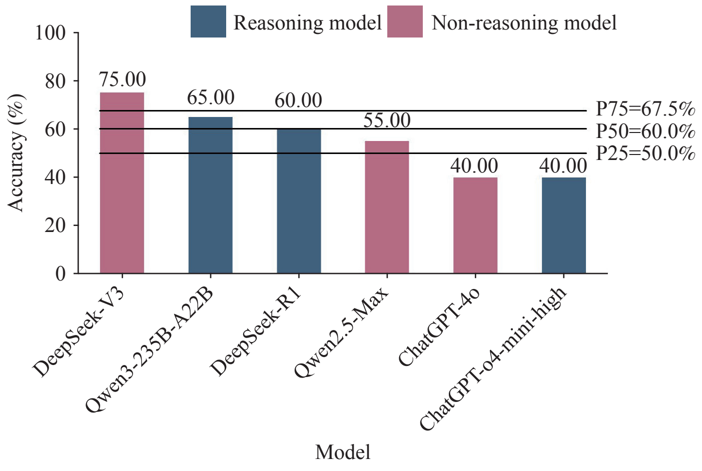


FIGURE 1. Accuracy of each model in multiple-choice questions.

accuracy rate of junior epidemiologists (60.0%) (*P*<0.05).

Table 1 demonstrates strong inter-rater reliability between the two experts in their evaluation of the case-based questions. Consequently, we utilized the average of both evaluators' scores as the final assessment for each open-ended question.

In the case-based section, performance varied across questions and models. For Question 1, DeepSeek-V3 achieved the highest score and was the only model to exceed the 75th percentile (P75) of junior epidemiologist scores. For Question 2, ChatGPT-4o demonstrated superior performance, while Qwen2.5 and DeepSeek-R1 both matched the P75 level of junior epidemiologists. For Question 3, four models — Qwen2.5, Qwen3-235B-A22B, DeepSeek-V3, and DeepSeek-R1 — all scored above the P75 level of junior epidemiologists, with Qwen2.5 achieving the highest score. For Question 4, all models except ChatGPT-o4-mini-high exceeded the P75 level of junior epidemiologists, with ChatGPT-4o demonstrating the strongest performance.

The chi-squared value from the Friedman test was 6.765, with a *P* of 0.239. Paired Wilcoxon tests revealed that the differences between ChatGPT-o4-mini-high and the other five models (DeepSeek-R1, *P*=0.11; DeepSeek-V3, *P*=0.11; Qwen3-235B-A22B, *P*=0.11, Qwen2.5-Max, *P*=0.10, ChatGPT-4o, *P*=0.34) were not statistically significant. All other pairwise comparisons yielded *P* greater than 0.5.

TABLE 1. Correlation between scores assigned by two evaluators for responses provided by six large language models.

Question	Pearson correlation		Spearman correlation	
	r	*P*	*ρ*	*P*
Question 1	0.937	0.006	0.742	0.091
Question 2	0.859	0.028	0.857	0.029
Question 3	0.860	0.028	0.739	0.094
Question 4	0.970	0.001	0.953	0.003

DISCUSSION

This study evaluated the capabilities of six currently popular LLMs in supporting field epidemiology investigations and compared their performance with examination scores from junior field epidemiologists. Among the multiple-choice questions, DeepSeek-V3 achieved the highest accuracy rate, followed by Qwen3-235B-A22B and DeepSeek-R1. For the case-based questions, no statistically significant differences were observed among the models overall; however, ChatGPT-o4-mini-high demonstrated relatively poor performance compared to the other models.

In this study, the Chinese-language LLMs (DeepSeek and Qwen) demonstrated superior performance compared to ChatGPT. The DeepSeek and Qwen models were developed using extensive Chinese language corpora during training, whereas ChatGPT was trained with limited Chinese-language content (*4*). Consequently, ChatGPT performed poorly on questions that relied heavily on Chinese language knowledge or cultural context. However, for tasks such as data analysis (Question 4), which are less dependent on Chinese-language training data, ChatGPT exhibited acceptable performance.

This study revealed that, for multiple-choice questions, most LLMs achieved lower accuracy rates than the 75th percentile level of junior field epidemiologists. Conversely, in the case-based

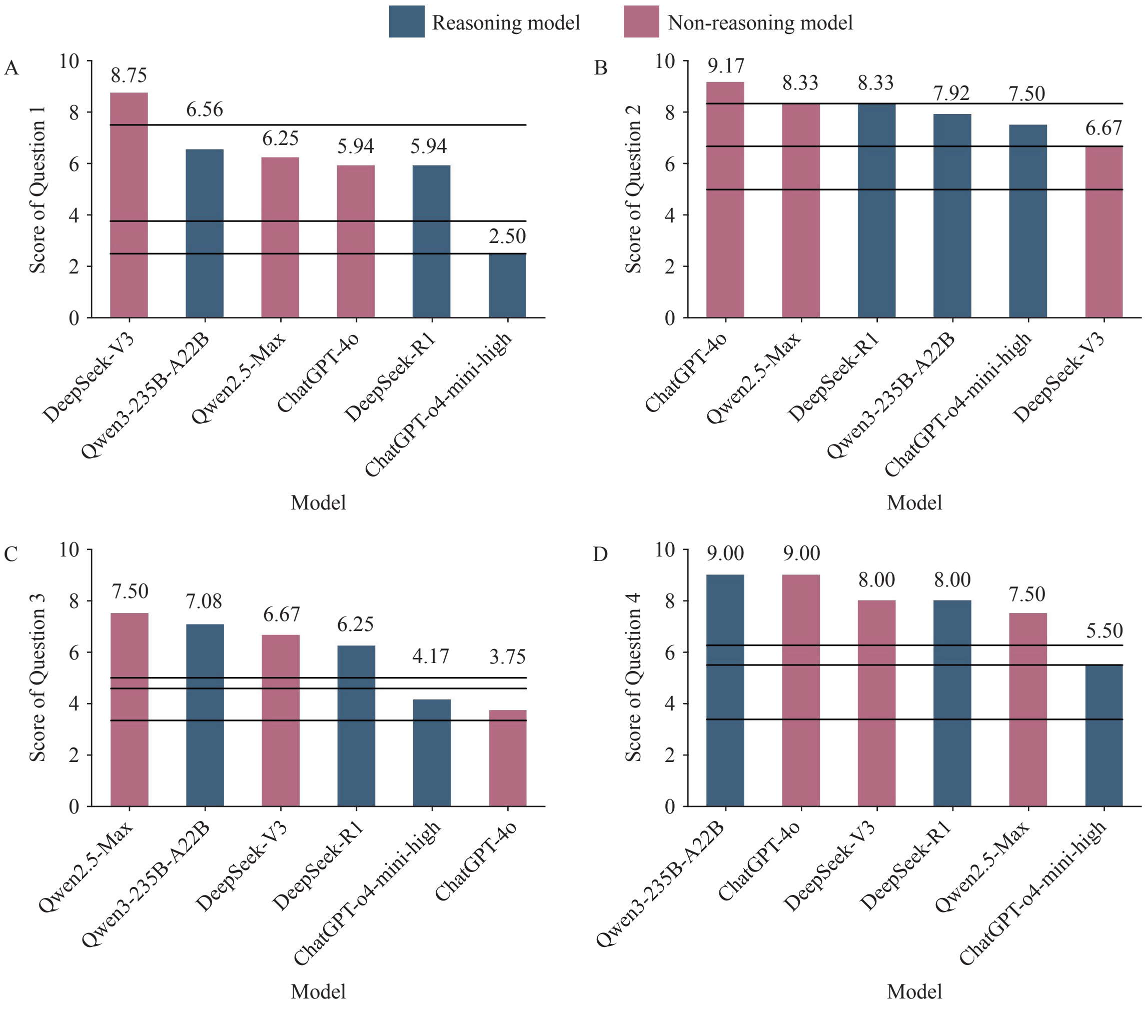


FIGURE 2. The average scores for the answers to each open-ended question provided by the six large language models. (A) Average Scores for Question 1; (B) Average Scores for Question 2; (C) Average Scores for Question 3; (D) Average Scores for Question 4.

questions, the overall performance of LLMs exceeded that of most junior field epidemiologists. Poor performance was particularly evident on Question 1, which involved professional prevention and control protocols for specific infectious diseases. This limitation may primarily stem from the absence of specialized knowledge resources in the LLMs (*4*). Similarly, the LLMs scored relatively low on Question 3, which addressed outbreak control measures. Their responses frequently included irrelevant or non-essential content, likely due to the same knowledge gap, resulting in answers that lacked technical precision and professional rigor.

Previous research has indicated that closed-source models may outperform their open-source counterparts (*5*). However, our findings demonstrate that the four Chinese open-source models generally exceeded ChatGPT's performance, underscoring the substantial potential of open-source architectures. Open-source models offer the advantage of local deployment, providing enhanced data security — a feature of paramount importance for developing specialized LLMs tailored to public health institutions.

Our study also revealed that reasoning models did not demonstrate superior performance compared to non-reasoning models, a finding consistent with observations by Sandmann et al. (*6*). Through chain-of-thought prompting in the reasoning models, we observed that LLMs incorporate knowledge from various temporal periods within their training datasets. However, these models lack the capability to distinguish between outdated and current information, resulting in instances where they failed to provide the most up-to-date knowledge.

Nevertheless, the implementation of LLMs in field epidemiology investigations continues to face several significant challenges. A critical concern is that field epidemiology is intrinsically linked to disease prevention and control, which demands exceptional timeliness and accuracy in model outputs. Our investigation identified limitations regarding citation accuracy in LLM-generated responses. In the case-based questions, several LLMs referenced guidelines or technical documents that were entirely fabricated. This presents substantial risks for junior professionals who may depend on these models without possessing the expertise to identify such erroneous references. Furthermore, LLMs trained on public knowledge bases carry an inherent risk of data contamination, potentially compromising the reliability of their outputs. These limitations have been documented in the existing literature (*7*–*8*). We therefore strongly recommend that professionals exercise caution when utilizing LLMs, cross-reference their outputs against established trusted sources, and treat these models as supplementary tools rather than substitutes for individual knowledge and experience. To enhance model performance, developing specialized knowledge resources for LLMs will be essential, supported by high-quality, regularly updated datasets for training purposes.

Another critical challenge involves data security and privacy protection (*9*). Field epidemiology investigations frequently handle sensitive information, including patient privacy data and confidential government decision-making processes, all requiring robust protection measures. Without adequate safeguards, the practical implementation of LLMs could face severe limitations. To address these concerns, comprehensive regulatory frameworks will play an essential role. The European Union has already established relevant regulations through the *EU AI Act*, representing the world's first comprehensive artificial intelligence legislation. In 2023, China also issued *China's Interim Measures for the Management of Generative AI Services*. However, as an emerging technology, LLM governance and oversight require continued research and development to ensure both innovation advancement and safety assurance (*10*).

This study presents several limitations that warrant consideration. First, our evaluation was restricted to entrance exam questions from the Zhejiang Field Epidemiology Training Program, which may not comprehensively represent all aspects of field epidemiology investigations. Second, LLM outputs exhibit inherent stochasticity, meaning responses to identical prompts may vary across individual runs. However, existing research indicates that for knowledge-intensive tasks, while model performance may show sensitivity to minor prompt variations, it generally maintains relative stability overall. Finally, our evaluation employed a limited number of questions, with case-based scenarios focusing exclusively on infectious diseases. Consequently, model performance in other types of public health emergencies remains uncertain. Future studies should expand the evaluation scope to enhance the reliability and generalizability of these findings.

This study evaluated the potential of six leading LLMs to support field epidemiology investigations by comparing their performance against junior field epidemiologists' examination scores. Our findings demonstrate that several models achieved notable accuracy and relevance across both multiple-choice and case-based assessments. However, current LLMs cannot yet replace human epidemiological expertise. While these models show promise as supplementary tools, their practical implementation faces significant challenges. Future development should prioritize integrating verified knowledge databases to optimize model performance and establishing robust regulatory frameworks to ensure their safe and effective application in public health settings.

Conflicts of interest: No conflicts of interest.

Funding: Supported by the Medical Health Science and Technology Project of Zhejiang Provincial Health Commission (2024KY896), the Public Health Talent Development Support Program of National Disease Control and Prevention Administration (2025) and the Zhejiang Provincial Key Laboratory of Vaccine and Preventive Research on Infectious Diseases from Department of Science and Technology of Zhejiang Province (2024).

doi: 10.46234/ccdcw2025.220

[#] Corresponding author: Chen Wu, chenwu@cdc.zj.cn.

[1] Department of Public Health Surveillance and Advisory, Zhejiang Provincial Center for Disease Control and Prevention, Hangzhou City, Zhejiang Province, China; [2] Department of Communicable Disease Control and Prevention, Shaoxing Center for Disease Control and Prevention, Shaoxing City, Zhejiang Province, China; [3] Department of Communicable Disease Control and Prevention, Zhoushan Center for Disease Control and Prevention, Zhoushan City, Zhejiang Province, China.

[&] Joint first authors.

Submitted: June 20, 2025
Accepted: September 27, 2025
Issued: October 10, 2025

REFERENCES

1. Conroy G, Mallapaty S. How China created AI model DeepSeek and shocked the world. Nature 2025;638(8050):300 – 1. https://doi.org/10.1038/d41586-025-00259-0.
2. Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary evidence of the use of generative AI in health care clinical services: systematic narrative review. JMIR Med Inform 2024;12:e52073. https://doi.org/10.2196/52073.
3. Rasmussen SA, Goodman RA. The CDC field epidemiology manual. New York: Oxford University Press. 2018. https://www.amazon.com/CDC-Field-Epidemiology-Manual/dp/0190624248.
4. Wu JG, Wu X, Qiu ZP, Li MH, Lin SX, Zhang YY, et al. Large language models leverage external knowledge to extend clinical insight beyond language boundaries. J Am Med Inform Assoc 2024;31(9):2054 – 64. https://doi.org/10.1093/jamia/ocae079.
5. Nazi ZA, Hossain R, Mamun FA. Evaluation of open and closed-source LLMs for low-resource language with zero-shot, few-shot, and chain-of-thought prompting. Natl Lang Process J 2025;10:100124. https://doi.org/10.1016/j.nlp.2024.100124.
6. Sandmann S, Hegselmann S, Fujarski M, Bickmann L, Wild B, Eils R, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med 2025;31(8):2546 – 9. https://doi.org/10.1038/s41591-025-03727-2.
7. Clelland CL, Moss S, Clelland JD. Warning: artificial intelligence chatbots can generate inaccurate medical and scientific information and references. Explor Digit Health Technol 2024;2:1 – 6. https://doi.org/10.37349/edht.2024.00006.
8. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare 2023;11(6):887. https://doi.org/10.3390/healthcare11060887.
9. The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digit Health 2023;5(3):e102. http://dx.doi.org/10.1016/s2589-7500(23)00023-7.
10. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digit Med 2023;6(1):120. https://doi.org/10.1038/s41746-023-00873-0.

SUPPLEMENTARY MATERIALS

Access Date and Settings for the Large Language Models

All LLMs were accessed through their web interfaces on May 12, 2025. The DeepSeek-V3 model was the 0324 build version, whereas specific build versions were not disclosed for DeepSeek-R1 (initial release), Qwen, or ChatGPT. No additional tools or plugins were used. All reasoning models displayed the chain of thought by default. Each model was queried using a newly registered account that had not been used for any prior interactions, ensuring no influence from historical usage or personalization. The prompts did not include chain-of-thought or “reasoning” related instructions. To ensure fairness, all chat memory and user personalization settings were disabled. This prevented models from benefiting from prior context and guaranteed that each query was processed independently. The specific settings were as follows:

Qwen: Switched to the test model with the temporary conversation setting enabled. Internet search was disabled; all other settings remained at their defaults.

DeepSeek: Switched to the test model. Internet search was disabled, with all other settings kept as defaults.

ChatGPT: Switched to the test model with the temporary conversation setting enabled. All other settings kept as defaults.

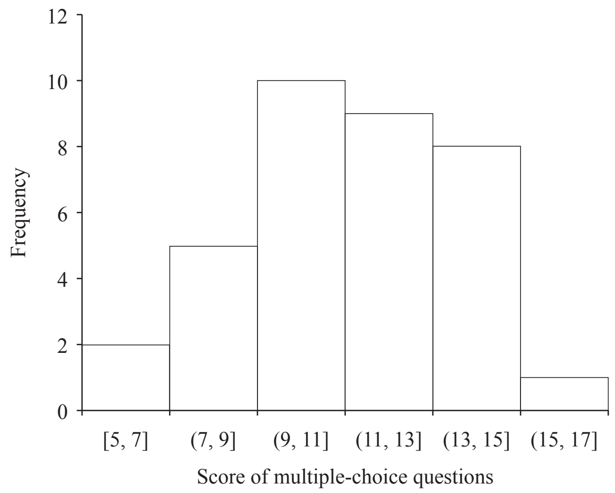
The prompts were as follows:

For case-based questions: “Please play the role of an on-site epidemiological investigation expert from the CDC and answer the following question”.

For multiple-choice questions: “Below is the written examination for the enrollment of a field epidemiology training program at a provincial CDC. Please provide your answers and indicate which ones you believe to be correct”.

Performance of the Junior Epidemiologists on the Multiple-Choice Questions

A total of 35 junior epidemiologists participated in the examination, with a mean score of 11.4 (standard deviation =2.6). The score distribution was as follows: minimum =5/20, 25th percentile (P25) =10/20, median (P50) =12/20, 75th percentile (P75) =13.5/20, and maximum =16/20.



SUPPLEMENTARY FIGURE S1. Histogram of score distribution for 35 junior epidemiologists who took the examination.

SUPPLEMENTARY TABLE S1. Performance of the six large language models on the multiple-choice questions.

Models	Accuracy (%)	95% confidence interval (%)	*P**
DeepSeek-V3	75.0	50.9, 91.3	<0.001
Qwen-235B-A22B	65.0	40.8, 84.6	<0.001
DeepSeek-R1	60.0	36.1, 80.9	<0.001
Qwen2.5-Max	55.0	31.5, 76.9	<0.001
ChatGPT-4o	40.0	19.1, 63.9	0.262
ChatGPT-o4-mini-high	40.0	19.1, 63.9	0.262

* Two-sided Bonferroni-adjusted *P* for comparing the model with chance.