https://www.sworldjournal.com/index.php/swj/article/view/swj35-02-006
DOI: 10.30888/2663-5712.2026-35-02-006

UDC: 61:378.244:004.89

PERFORMANCE OF CHATGPT-5.1 AND GEMINI 2.5 FLASH ON THE UKRAINIAN LICENSING INTEGRATED EXAMINATION "KROK 3": A COMPARATIVE STUDY

Yaroslav Mykhalko
c.m.s., as.prof.
ORCID: 0000-0002-9890-6665
Yaroslav Filak
c. pe & s.s, as.prof.
ORCID: 0000-0002-7510-263X
Iryna Filak
c.f.s., as.prof.
ORCID: 0000-0002-8573-4040
Felix Filak
c.m.s., as.prof.
ORCID: 0000-0001-7595-5416
Yelyzaveta Rubtsova
c.m.s., as.prof.
ORCID: 0000-0001-9395-1822
SHEI "Uzhhorod National University",
Uzhhorod, Narodna Square, 3, 88000

Abstract. *The development of large language models (LLMs) has opened new perspectives for their integration into healthcare. Previous studies have demonstrated the high effectiveness of leading LLMs, such as GPT-4, in passing standardized English-language medical examinations (e.g., the USMLE), showing accuracy levels comparable to those of qualified practicing physicians. However, data on the performance of these models in non-English, particularly Ukrainian, contexts remain limited.*

Aim. *To evaluate the performance of contemporary LLMs in completing test items of the Ukrainian-language licensing integrated examination "Krok 3" and to analyze the impact of role-based instructions on the accuracy and stability of responses.*

Materials and Methods. *The performance of ChatGPT-5.1 (OpenAI) and Gemini 2.5 Flash (Google) was assessed using a set of 150 official test items in the specialty "Internal Medicine" from the Ukrainian-language licensing integrated examination "Krok 3." The models were tested in a standard mode and in a mode with the role-based instruction "act as a professional physician." The percentage of correct answers, 95% confidence intervals (Wilson method), statistical differences (McNemar test), and agreement between modes (Cohen's weighted kappa, k_w) were evaluated to assess accuracy and response stability.*

Results. *Both models demonstrated high accuracy (>90%) in both operating modes. No statistically significant differences were found between the models or between modes with and without the role-based instruction ($p > 0.05$). The role-based instruction did not affect overall performance. However, response stability differed: ChatGPT showed moderate agreement between modes ($k_w = 0.41$), whereas Gemini demonstrated high agreement ($k_w = 0.80$).*

Conclusions. *Contemporary LLMs are capable of performing Ukrainian-language licensing examination tasks from "Krok 3" with high accuracy, indicating effective multilingual adaptation in the field of medical education. Role-based instructions do not improve response accuracy in the single-best-answer MCQ format, although response stability may vary between models. These*

findings highlight the substantial potential of LLMs for use in Ukrainian medical education, preparation for licensing examinations, and as clinical decision-support tools, provided that further local validation is conducted.

Key words: *large language models, artificial intelligence, medical licensing examination, Krok 3, multiple-choice questions, role-based prompting, response stability.*

Introduction.

The past decade has been characterized by the unprecedented development of artificial intelligence (AI) systems, with large language models (LLMs) occupying a central position in this progress. These models are typically built on the Transformer architecture and incorporate advanced algorithmic innovations, such as multimodality, which has led to a substantial increase in their performance across a wide range of tasks [1,2]. The scaling of model size and training datasets, as described by the so-called "scaling laws," has been a key factor in achieving state-of-the-art results [3].

At the early stages of integration into healthcare, LLMs were primarily focused on improving operational efficiency, including the automation of administrative processes, summarization of clinical notes, extraction and analysis of patient data, and diagnostic reports generation [1]. The emergence of new LLM versions and the continuous improvement of their performance have been accompanied by a shift in scientific interest to assessing the ability of these models to solve highly complex tasks requiring deep knowledge synthesis. To quantitatively evaluate the clinical competence and effectiveness of LLMs, the research community widely employs standardized medical licensing examinations, such as the United States Medical Licensing Examination (USMLE) [4,5]. The results of previous studies are highly compelling. For example, leading LLMs such as GPT-4 have achieved high accuracy rates, in some cases within the range of 80–90%, across all three steps of the USMLE, while DeepSeek achieved 89% accuracy on Step 1 [6,7]. Moreover, in an assessment analogous to the European Diploma in Intensive Care (EDIC) examination, GPT-4o demonstrated the highest performance (89.0%), significantly exceeding the average performance of human physicians (61.9%) [8]. Comparable results have been reported for other standardized English-language examinations, including IFOM and UKMLE [9]. These findings provide evidence that LLMs have reached a level at which they can be considered reliable sources of medical knowledge.

Despite these impressive results on English-language examinations, studies evaluating the performance of LLMs on non-English medical licensing exams remain relatively rare. Although some research has assessed LLM performance on the Chinese National Medical Licensing Examination (CNMLE), the German Medical Licensing Examination (GMLE), the Polish Medical Licensing Examination (LEK), and the Japanese Medical Licensing Examination (JMLE), these studies have already revealed substantial variability associated with linguistic and geographical factors [4,9-13]. For instance, GPT-4.0 achieved a passing score on the GMLE with an accuracy exceeding 70%, in some cases outperforming medical students [9]. This variability underscores that the multilingual adaptation of LLMs is not universal, and that transitioning to languages other than English often results in reduced overall performance.

The Ukrainian Licensing Integrated Examination “Krok 3” occupies a central position in the national medical licensing system. It is a mandatory examination designed to assess whether physicians completing internship training in a given specialty meet the required level of professional competence. To pass the examination, candidates must correctly answer at least 66% of the test items. This represents a critical stage in the certification process: interns who fail “Krok 3” are not admitted to subsequent stages of attestation and therefore do not receive certification as medical specialists. The “Krok 3” examination typically consists of 150 multiple-choice single-best-answer questions, with a total testing time of 150 minutes. It is administered across various medical specialties and has a strong clinical orientation, requiring candidates not only to establish diagnoses but also to determine treatment strategies, select appropriate investigations, and apply pharmacological therapies rationally.

Assessing the effectiveness of LLMs in passing the “Krok 3” examination is important because it enables evaluation of their ability to accurately reproduce professional medical competencies in the Ukrainian language, serving as an indicator of their adaptation to the linguistic, terminological, and contextual environment of national medical education. Testing LLMs in Ukrainian assesses their ability to handle localized medical terminology, data, and regulatory context, which is essential for educational and clinical use. Evaluating LLM performance in the context of the

Ukrainian-language "Krok 3" examination is therefore a necessary prerequisite for ensuring the safe, evidence-based, and scientifically valid implementation of these technologies in Ukrainian medical education.

Aim. To evaluate the performance of contemporary LLMs in completing test items of the Ukrainian-language licensing integrated examination "Krok 3" and to analyze the impact of role-based instructions on the accuracy and stability of responses.

Materials and Methods. The study assessed the performance of ChatGPT-5.1 (OpenAI) and Gemini 2.5 Flash (Google) in solving standardized test items from the "Krok 3" examination. The LLMs were used in a standard operating mode without activation of specialized reasoning features, deep research modes, external tools, or any other advanced capabilities. All responses were generated within ordinary text-based user-model interactions.

A total of 150 test items from the official database (https://test.testcentr.org.ua/) in the specialty "Internal Medicine" were included. All questions were formatted as single-best-answer multiple-choice items and did not contain images. Correct answers were determined according to the official answer key.

The study was conducted in December 2025. Each model was tested in two operating modes: Standard Mode (SM) and Clinician-Instructed Mode (CIM). In both modes, models were presented with the complete set of 150 questions and instructed to select the correct answer from the provided options; however, in CIM, models received an additional instruction to act as a professional physician. In all cases, the order of questions was identical, and only the selected answer option was recorded, without any additional explanations. To avoid the influence of prior context, each mode was evaluated in a separate chat session, and the memory function was disabled.

Model performance was assessed based on the percentage of correct answers. 95% confidence intervals (CIs) were calculated using the Wilson method. Statistical significance of differences between SM and CIM, as well as between models within the same mode, was evaluated using a two-sided McNemar test. A p value < 0.05 was considered statistically significant. To assess the stability of answer selection between the two operating modes for each model, weighted Cohen's kappa (k_w) was calculated

and interpreted according to the following scale: <0.0, poor; 0.0-0.2, slight; 0.2-0.4, fair; 0.4-0.6, moderate; 0.6-0.8, substantial; and 0.8-1.0, almost perfect agreement [14].

The study used only anonymized test items and did not involve personal data or human participants. The experiment was conducted exclusively with language models; therefore, no additional ethical approval was required.

Results. In Standard Mode (SM), both models demonstrated high response accuracy exceeding 90% (Table 1). Although Gemini 2.5 Flash performance was slightly higher comparing to ChatGPT-5.1, the paired McNemar test revealed no statistically significant differences between the models ($p > 0.05$).

Table 1 – Performance of ChatGPT-5.1 and Gemini 2.5 Flash on test items in the specialty "Internal Medicine" under different operating modes.

LLM / mode	Correct answers, n (%)	95% CI	k_w (95% CI)
ChatGPT-SM	138 (92.00)	86,54-95,36	0,41 (0.16-0.66)
ChatGPT-CIM	136 (90.67)	84,94-94,36	
Gemini-SM	139 (92.67)	87,35-95,86	0,80 (0.62-0.99)
Gemini-CIM	139 (92.67)	87,35-95,86	

Source: compiled by the authors of this study

After applying the role-based instruction "act as a professional physician," the accuracy of ChatGPT decreased by 1.33%; however, this difference was not statistically significant compared with the results obtained in the previous mode ($p > 0.05$).

In the case of Gemini, performance remained identical to that observed in SM and was slightly higher than that of ChatGPT-5.1 in the same mode. Nevertheless, this difference was also not statistically significant ($p > 0.05$). These data indicate that the role-based instruction did not affect the overall success rate of either model.

Analysis of response agreement between modes revealed differences in model stability. ChatGPT exhibited moderate agreement ($k_w = 0.41$), suggesting that some individual responses changed when transitioning from SM to Clinician-Instructed Mode (CIM), even though overall accuracy remained similar. In contrast, Gemini

showed high agreement between the two modes (k_w = 0.80), indicating stable answer selection regardless of instruction phrasing. Despite the slight numerical advantage of Gemini 2.5 Flash in accuracy, comparisons between the two models across different modes confirmed their equivalence ($p > 0.05$). The 95% CIs for all four conditions overlapped substantially, further underscoring the absence of clinically and statistically significant differences. The overall high proportion of correct answers demonstrates the ability of both models to effectively handle Ukrainian-language textual test items in the specialty "Internal Medicine."

Discussion. Despite their global potential, the implementation of LLMs in clinical practice requires rigorous validation to ensure accuracy and patient safety [1]. The success of LLMs on high-stakes licensing examinations suggests that these models possess knowledge comparable to that of qualified practicing physicians.

At the same time, there is a risk that such models may reflect and amplify existing societal biases, including racial or gender, due to imbalances in the data used for training, potentially leading to biased diagnostic outputs [15]. It has been demonstrated that LLMs may generate inconsistent or inaccurate responses in non-English contexts, particularly when referring to clinical guidelines [16]. This underscores the need for local verification to mitigate disparities in healthcare delivery.

Language-specific factors pose unique linguistic and methodological challenges for global LLMs, as they are predominantly trained on English-language or other high-resource corpora. In the case of the Ukrainian language, the challenge extends beyond simple translation. Ukraine is currently undergoing a process of final standardization and unification of medical terminology in professional communication [17]. Although medical terminology constitutes a coherent system, the presence of terminological variation and the need to restore national medical lexicon complicate the task for LLMs. A model must not only "know" a medical fact but also formulate it correctly using normative or accepted terms that may differ from those encountered during training [17].

Evaluating LLM performance in the context of the "Krok 3" examination represents a direct test of their potential professional applicability within the national

regulatory and linguistic environment. Such research is not only a scientific opportunity but also an ethical imperative to ensure the development of fair and safe AI solutions in the Ukrainian healthcare system.

Although both models in the present study were presented with "Krok 3" examination items in Ukrainian, they demonstrated high overall performance comparable to that reported for English-language examinations such as the USMLE [18]. This indicates not only their ability to correctly interpret medical information but also their apparent "understanding" of the nuances of the Ukrainian language in general and Ukrainian medical terminology in particular.

The absence of statistically significant differences between models ($p > 0.05$) and between modes with and without role-based instructions confirms that stylistic modifications of prompts do not influence the final answer selection in standardized clinical tasks. At the same time, the combination of high baseline performance and the lack of an effect from role-based instructions may be explained by the fact that the models already "know" the correct answers for most typical clinical scenarios in internal medicine. The observed difference in response stability (k_w = 0.41 and 0.80 ChatGPT and Gemini respectively) indicates greater variability in ChatGPT's responses when instructions are modified, suggesting a higher internal sensitivity of this model to prompt formulation. Moreover, although the order of questions was identical, differences in ChatGPT responses may reflect a cumulative contextual effect within a chat session, influencing partial answer changes despite similar overall accuracy. In contrast, Gemini appears to employ a more deterministic answer-selection mechanism, resulting in high agreement across modes.

Several limitations of this study should be acknowledged. First, the analysis was restricted to test items from a single specialty ("Internal Medicine"), limiting direct extrapolation of the results to other medical disciplines or to "Krok 3" examinations in other specialties. Second, all items were exclusively text-based and did not include images. Third, the fixed order of question presentation may have contributed to context accumulation within individual sessions, potentially affecting answer selection. Fourth, the evaluation was conducted using specific model versions available at the time of

testing; given the rapid and ongoing updates of LLMs, the results may not fully reflect future performance. Finally, correct answers were determined based on the existing test database without additional expert clinical verification, which may have influenced the interpretation of certain ambiguous cases.

Conclusions. The results confirm that contemporary LLMs demonstrate a high level of accuracy in completing text-based internal medicine test items within the context of the Ukrainian licensing examination "Krok 3." The use of role-based instructions ("act as a physician") does not have a statistically significant or clinically meaningful impact on model performance in the single-best-answer MCQ format, although response stability may differ between systems. These findings support the effectiveness of LLMs in standardized medical testing and highlight their potential for use in medical education, licensing examination preparation, and simulation-based learning environments, provided that further local validation is conducted.

References

1. Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M., & Božić, J. (2025). A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. Healthcare, 13(6), 603. https://doi.org/10.3390/healthcare13060603.

2. Advances in Large Language Models for Medicine – arXiv, https://arxiv.org/html/2509.18690v1.

3. Lin, C., & Kuo, C.-F. (2025). Roles and Potential of Large Language Models in Healthcare: A Comprehensive Review. *Biomedical Journal*, 100868. https://doi.org/10.1016/j.bj.2025.100868.

4. Zong, H., Wu, R., Cha, J., Wang, J., Wu, E., Li, J., Zhou, Y., Zhang, C., Feng, W., & Shen, B. (2024). Large Language Models in Worldwide Medical Exams: Platform Development and Comprehensive Analysis. Journal of medical Internet research, 26, e66114. https://doi.org/10.2196/66114.

5. Nouri, H., Mahdavi, A., Abedi, A., Mohammadnia, A., Hamedan, M., & Amanzadeh, M. (2025). Performance of large language models in medical licensing

examinations: a systematic review and meta-analysis. Journal of educational evaluation for health professions, 22, 36. https://doi.org/10.3352/jeehp.2025.22.36.

6. Yang, Z., Yao, Z., Tasmin, M., Vashisht, P., Jang, W. S., Ouyang, F., Wang, B., McManus, D., Berlowitz, D., & Yu, H. (2025). Unveiling GPT-4V's hidden challenges behind high accuracy on USMLE questions: Observational Study. Journal of medical Internet research, 27, e65146. https://doi.org/10.2196/65146.

7. Nori, H., King, N., McKinney, S.M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on Medical Challenge Problems. ArXiv, abs/2303.13375.

8. Workum, J. D., Volkers, B. W. S., van de Sande, D., Arora, S., Goeijenbier, M., Gommers, D., & van Genderen, M. E. (2025). Comparative evaluation and performance of large language models on expert level critical care questions: a benchmark study. Critical care (London, England), 29(1), 72. https://doi.org/10.1186/s13054-025-05302-0.

9. Kasagga, A., Sapkota, A., Changaramkumarath, G., Abucha, J. M., Wollel, M. M., Somannagari, N., Husami, M. Y., Hailu, K. T., & Kasagga, E. (2025). Performance of ChatGPT and Large Language Models on Medical Licensing Exams Worldwide: A Systematic Review and Network Meta-Analysis With Meta-Regression. Cureus, 17(10), e94300. https://doi.org/10.7759/cureus.94300.

10. Wu, J., Wang, Z., & Qin, Y. (2025). Performance of DeepSeek-R1 and ChatGPT-4o on the Chinese National Medical Licensing Examination: A Comparative Study. Journal of medical systems, 49(1), 74. https://doi.org/10.1007/s10916-025-02213-z.

11. Guillen-Grima, F., Guillen-Aguinaga, S., Guillen-Aguinaga, L., Alas-Brun, R., Onambele, L., Ortega, W., Montejo, R., Aguinaga-Ontoso, E., Barach, P., & Aguinaga-Ontoso, I. (2023). Evaluating the Efficacy of ChatGPT in Navigating the Spanish Medical Residency Entrance Examination (MIR): Promising Horizons for AI in Clinical Medicine. Clinics and practice, 13(6), 1460–1487. https://doi.org/10.3390/clinpract13060130.

12. Fujimoto, M., Kuroda, H., Katayama, T., Yamaguchi, A., Katagiri, N., Kagawa, K., Tsukimoto, S., Nakano, A., Imaizumi, U., Sato-Boku, A., Kishimoto, N.,

Itamiya, T., Kido, K., & Sanuki, T. (2024). Evaluating Large Language Models in Dental Anesthesiology: A Comparative Analysis of ChatGPT-4, Claude 3 Opus, and Gemini 1.0 on the Japanese Dental Society of Anesthesiology Board Certification Exam. Cureus, 16(9), e70302. https://doi.org/10.7759/cureus.70302.

13. Liu, M., Okuhara, T., Dai, Z., Huang, W., Gu, L., Okada, H., Furukawa, E., & Kiuchi, T. (2025). Evaluating the Effectiveness of advanced large language models in medical Knowledge: A Comparative study using Japanese national medical examination. International journal of medical informatics, 193, 105673. https://doi.org/10.1016/j.ijmedinf.2024.105673.

14. Gwet K. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. Oxford: Advanced Analytics, LLC, Gaithersburg; 2014, p. 62-65.

15. Yang, Y., Jin, Q., Zhu, Q., Wang, Z., Erramuspe Álvarez, F., Wan, N., Hou, B., & Lu, Z. (2025). Beyond Multiple-Choice Accuracy: Real-World Challenges of Implementing Large Language Models in Healthcare. Annual Review of Biomedical Data Science. https://doi.org/10.1146/annurev-biodatasci-103123-094851.

16. Schlicht, I.B., Zhao, Z., Sayin, B., Flek, L., Rosso, P. (2025). Do LLMs Provide Consistent Answers to Health-Related Questions Across Languages?. In: Hauff, C., et al. Advances in Information Retrieval. ECIR 2025. Lecture Notes in Computer Science, vol 15574. Springer, Cham. https://doi.org/10.1007/978-3-031-88714-7_30.

17. Vasylovska, I. (2017). The modern state and problems of ukrainian medical terminography. Theory and Practice of Teaching Ukrainian as a Foreign Language, (13), 116–121.

18. Bicknell, B. T., Butler, D., Whalen, S., Ricks, J., Dixon, C. J., Clark, A. B., Spaedy, O., Skelton, A., Edupuganti, N., Dzubinski, L., Tate, H., Dyess, G., Lindeman, B., & Lehmann, L. S. (2024). Critical Analysis of ChatGPT 4 Omni in USMLE Disciplines, Clinical Clerkships, and Clinical Skills (Preprint). JMIR Medical Education. https://doi.org/10.2196/63430.

Article sent: 01.10.2022