








## ISSUES OF EDUCATION IN ANESTHESIOLOGY AND CRITICAL CARE MEDICINE








## ВОПРОСЫ ОБРАЗОВАНИЯ В АНЕСТЕЗИОЛОГИИ И РЕАНИМАТОЛОГИИ

<https://doi.org/10.21320/1818-474X-2026-2-176-185>

### Using large language models for solving tests in anesthesiology and intensive care: a comparative study

### Использование больших языковых моделей для решения тестовых заданий по анестезиологии и реаниматологии: сравнительное исследование

A.A. Klimov <sup>1,\*</sup>, A.V. Karelin <sup>1</sup>, S.B. Liapustin <sup>2</sup>, S.I. Rudnitsky <sup>1</sup>, M.A. Tolstova <sup>1</sup>, A.E. Shamonin <sup>1</sup>, V.V. Subbotin <sup>1,3</sup>

А.А. Климов <sup>1,\*</sup>, А.В. Карелин <sup>1</sup>, С.Б. Ляпустин <sup>2</sup>, С.И. Рудницкий <sup>1</sup>, М.А. Толстова <sup>1</sup>, А.Е. Шамонин <sup>1</sup>, В.В. Субботин <sup>1,3</sup>

<sup>1</sup> Loginov Moscow Clinical Scientific Center, Moscow, Russia

<sup>2</sup> E.A. Vagner Perm State Medical University, Perm, Russia

<sup>3</sup> Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology, Moscow, Russia

<sup>1</sup> ГБУЗ «Московский клинический научный центр им. А.С. Логинова» Департамента здравоохранения города Москвы, Москва, Россия

<sup>2</sup> ФГБОУ ВО «Пермский государственный медицинский университет им. академика Е.А. Вагнера» Минздрава России, Пермь, Россия

<sup>3</sup> ФГБНУ «Федеральный научно-клинический центр реаниматологии и реабилитологии» Минобрнауки России, Москва, Россия

#### Abstract

#### Реферат

**INTRODUCTION:** In recent years, large language models (LLMs) have found widespread application in healthcare. However, their effectiveness in solving specialized tests in Russian, particularly in anesthesiology and critical care, remains poorly studied. **OBJECTIVE:** To evaluate the performance of LLMs on single-answer multiple-choice questions in Russian anesthesiology and critical care, compared to results of resident physician teams from the “Professionals” competition at the Forum of Anaesthesiologists and Reanimatologists-2025. **MATERIALS AND METHODS:** We conducted a comparative study of responses to 30 test items from the qualifying stage of the “Professionals” competition. Results from 38 resident teams were compared against answers from LLMs: Generative Pre-trained Transformer (GPT)-4o, GPT-5, Alisa AI, DeepSeek V-3.2, GigaChat, Gemini 2.5 Flash, and Qwen3-Max. Comparison methods included rank analysis, pairwise comparison (win-rate), agreement assessment (Cohen’s  $\kappa$  coefficient), and correlation analysis ( $\phi$ -coefficient). **RESULTS:** The median score of participating teams was 24.5 out of 30, with one team achieving the maximum score (30/30). Four models (GPT-4o, GPT-5, DeepSeek V-3.2, Gemini 2.5 Flash) demonstrated 100 % accuracy (30 points), sharing first rank position with the leading team. These models achieved the

**АКТУАЛЬНОСТЬ:** В последние годы большие языковые модели (Large Language Models — LLM) нашли широкое применение в сфере здравоохранения. Эффективность данных моделей при решении узкоспециализированных тестовых заданий на русском языке, в частности по анестезиологии и реаниматологии, остается малоизученной. **ЦЕЛЬ ИССЛЕДОВАНИЯ:** Оценить успешность выполнения тестовых заданий с одним вариантом правильного ответа на русском языке по специальности «Анестезиология и реаниматология» современными LLM (Generative Pre-trained Transformer [GPT]-4o, GPT-5, Alisa AI, DeepSeek V-3.2, GigaChat, Gemini 2.5 Flash, Qwen3-Max) в сравнении с результатами команд врачей-ординаторов, участвовавших в конкурсе «Профессионалы» на Форуме анестезиологов-реаниматологов России 2025 г. **МАТЕРИАЛЫ И МЕТОДЫ:** Сравнительное исследование ответов на 30 тестовых заданий отборочного этапа конкурса «Профессионалы». Результаты 38 команд врачей-ординаторов сопоставлены с ответами LLM: GPT-4o, GPT-5, Alisa AI, DeepSeek V-3.2, GigaChat, Gemini 2.5 Flash и Qwen3-Max. Для сравнения использовали ранговый анализ, попарное сравнение (pairwise win-rate), оценку согласованности (коэффициент  $\kappa$  Коэна) и корреляцион-

97th percentile, reflecting superiority over 37 of 38 participating teams. Qwen3-Max and Alisa AI scored 29.9 and 29 points respectively, ranking first and second in the overall rating (97th and 92nd percentiles). GigaChat provided no answers. The win-rate of LLMs against a randomly selected team varied from 0.97 to 1.00. Near-perfect agreement was observed among leading models ( $\kappa = 1.00$ ), with very high correlation between their answers and the majority choice of residents ( $\phi \approx 1.00$ ). No statistically significant differences were found between LLM results ( $p > 0.05$ ). **CONCLUSIONS:** Modern large language models demonstrate high accuracy in solving standardized tests in anesthesiology and critical care in Russian, significantly exceeding the median performance of resident physician teams.

**KEYWORDS:** anesthesiology, critical care, large language models, artificial intelligence, education, medical, medical tests

\* *For correspondence:* Andrei A. Klimov — Ph. D. of Medical Sciences, Anaesthesiologist of the Department of anesthesiology and resuscitation, Loginov Moscow Clinical Scientific Center, Moscow, Russia; e-mail: KlimovAA1@zdrav.mos.ru

✉ *For citation:* Klimov A.A., Karelin A.V., Liapustin S.B., Rudnitsky S.I., Tolstova M.A., Shamonin A.E., Subbotin V.V. Using large language models for solving tests in anesthesiology and intensive care: a comparative study. *Annals of Critical Care*. 2026; 2:176–185. <https://doi.org/10.21320/1818-474X-2026-2-176-185>

📅 *Received:* 03.12.2025

📅 *Accepted:* 15.02.2026

ный анализ ( $\phi$ -коэффициент). **РЕЗУЛЬТАТЫ:** Медианный результат команд-участников составил 24,5 балла из 30, максимальный результат (30 баллов из 30) набрала одна команда. Четыре модели (GPT-4o, GPT-5, DeepSeek V-3.2, Gemini 2.5 Flash) продемонстрировали 100%-я точность (30 баллов) при подготовке ответов, разделив 1-е ранговое место с командой-лидером. Процентиль данных моделей составил 97 %, что отражает их превосходство над 37 из 38 команд-участниц. Модели Qwen3-Max и Alisa AI показали результаты в 29,9 и 29 баллов соответственно, заняв 1-е и 2-е места в общем рейтинге (перцентили 97 и 92 %). Модель GigaChat не предоставила ответы. Доля «победы» LLM над случайно выбранной командой варьировала от 0,97 до 1,00. Наблюдалась почти полная согласованность ответов между моделями-лидерами ( $\kappa = 1,00$ ) и очень высокая корреляция их ответов с выбором большинства ординаторов ( $\phi \approx 1,00$ ). Статистически значимых различий между результатами LLM не выявлено ( $p > 0,05$ ). **ВЫВОДЫ:** Современные большие языковые модели демонстрируют высокую точность при решении стандартизированных тестов по анестезиологии и реаниматологии на русском языке, существенно превосходя медианные показатели команд врачей-ординаторов.

**КЛЮЧЕВЫЕ СЛОВА:** анестезиология, интенсивная терапия, большие языковые модели, искусственный интеллект, медицинское образование, тесты по медицине

\* *Для корреспонденции:* Климов Андрей Андреевич — канд. мед. наук, врач — анестезиолог-реаниматолог центра анестезиологии реаниматологии ГБУЗ «Московский клинический научный центр им А.С. Логинова» ДЗМ, Москва, Россия; e-mail: KlimovAA1@zdrav.mos.ru

✉ *Для цитирования:* Климов А.А., Карелин А.В., Ляпустин С.Б., Рудницкий С.И., Толстова М.А., Шамонин А.Е., Субботин В.В. Использование больших языковых моделей для решения тестовых заданий по анестезиологии и реаниматологии: сравнительное исследование. *Вестник интенсивной терапии им. А.И. Салтанова*. 2026; 2:176–185. <https://doi.org/10.21320/1818-474X-2026-2-176-185>

📅 *Поступила:* 03.12.2025

📅 *Принята к печати:* 15.02.2026

DOI: 10.21320/1818-474X-2026-2-176-185

## Introduction

In recent years, large language models (LLMs) have gained widespread adoption globally, representing a significant milestone in the evolution of digital technologies.

LLMs are artificial intelligence (AI) systems trained on vast amounts of text data, enabling them to comprehend the meaning and context of human language. These systems are based on the transformer architecture, introduced in 2017 [1], which utilizes an attention mechanism to analyze en-

tire texts and generate logical, coherent responses. In 2018, the first model of the Generative Pre-trained Transformer (GPT) series was developed using this architecture [2].

Publications exploring the potential applications of LLMs in healthcare began to emerge in the early 2020s. Even at that stage, researchers recognized the promising role of this AI paradigm in medicine [3, 4]. Currently, LLMs are being employed for various tasks, including the analysis of medical records, the automatic interpretation of laboratory and instrumental data with the generation of structured reports, and support for clinical decision-making. Furthermore, AI models demonstrate high accuracy in analyzing large medical datasets, such as electronic health records and laboratory results, paving the way for their integration into clinical information systems [5, 6]. Of particular interest is assessing the capabilities of AI within the specialty of anesthesiology and intensive care, a field where rapid information processing, time-sensitive decision-making, and effective team coordination are paramount.

A distinct area of LLM application is medical education and specialist training [7]. Models like ChatGPT and their analogs are used as interactive tools to explain complex concepts, analyze clinical scenarios, and train clinical thinking skills. The use of such systems enhances learning effectiveness, allowing students and residents to independently analyze errors and receive instant explanations for questions [7, 8]. LLMs are also employed to simulate exam scenarios and test knowledge in major medical disciplines, including anatomy, pharmacology, and pathophysiology [8, 9].

For an objective assessment of the level of knowledge and clinical thinking in this field, both in Russia and abroad, test items with a single correct answer are traditionally used, along with clinical case simulations that mirror real-life medical care situations [8–10]. Similar formats are used in examinations by the American Board of Anesthesiology (ABA), the European Society of Anaesthesiology and Intensive Care (ESAIC) for the European Diploma in Anaesthesiology and Intensive Care (EDAIC), as well as in the system for accreditation and final certification of physicians in the Russian Federation [11–13]. Studies analyzing the results of the United States Medical Licensing Examination (USMLE) and similar certification tests have shown that the performance of modern large language models is comparable to, and in some cases even exceeds, that of resident physicians and practicing doctors [8, 10].

In October 2025, within the framework of the annual Forum of Anesthesiologists and Intensive Care Specialists of Russia (FARR-2025), the “Professionals” competition for residents in anesthesiology and intensive care was held for the third time. It was attended by 38 teams from different regions of the Russian Federation, consisting of first- and second-year residents. The competition included solving test tasks at the qualifying stage, as well as performing specialty tasks and clinical situational problems designed to closely mimic real practice in an in-person quiz format.

After the competition, the organizers became interested in comparing the participants’ results with the answers that could be provided by modern large language models (GPT-4o, GPT-5, Alisa AI, DeepSeek V-3.2, GigaChat, Gemini 2.5 Flash). This idea arose from a series of observations demonstrating that many LLMs already show a high level of success in solving medical test tasks. However, the question remains: are such models capable of adequately answering highly specialized questions? At the same time, the effectiveness of using large language models for solving medical tasks in Russian remains virtually unexplored. Most published studies have been conducted in English, which prevents the direct extrapolation of their results to the conditions of clinical practice in the Russian Federation.

## Objective

To assess the performance of modern large language models (GPT-4o, GPT-5, Alisa AI, DeepSeek V-3.2, GigaChat, Gemini 2.5 Flash, Qwen3-Max) in solving single-answer multiple-choice questions in Russian within the specialty of anesthesiology and intensive care. This performance was compared with the results achieved by teams of resident physicians that took part in the “Professionals” competition, held at the 2025 Forum of Anesthesiologists and Intensive Care Specialists of Russia.

## Materials and methods

### Study design

A comparative study was conducted on the responses to test items from the qualifying stage of the 2025 “Professionals” competition in the specialty “Anesthesiology and Intensive Care”. The performance of the actual competition participants and modern large language models was evaluated. The study included the analysis of answers to 30 single-answer multiple-choice questions covering pathophysiology, diagnosis, and treatment of emergency conditions, legal aspects of emergency care, and interpretation of laboratory and instrumental data.

The study design was not reviewed by an ethics committee, as the work is based on the analysis of open data — comparing the results of large language models with the answers of participants in an educational competition. This analysis did not involve the use of personal data or any intervention in the process of providing medical care.

### Test materials

This study utilized a 30-question test from the qualifying stage of the “Professionals” competition. The test was developed by the competition’s methodological commit-

tee in accordance with the requirements for the theoretical training of residents in the specialty of anesthesiology and intensive care.

### Study participants

The actual study participants were first- or second-year clinical residents specializing in anesthesiology and intensive care who were registered for the competition. The analysis included the test results of 38 teams, each consisting of four individuals.

The study also incorporated test responses from the following large language models: GPT-4o (OpenAI), GPT-5 (OpenAI), Alisa AI (Yandex), DeepSeek V-3.2 (DeepSeek), GigaChat (Sberbank), Gemini 2.5 Flash (Google), and Qwen3-Max (Alibaba Group).

### Methods

On September 15, 2025, online testing for the participants of the “Professionals” competition was conducted on the MTS LINK platform. The testing was organized under conditions designed to closely mimic an examination setting: an online format, a time limit (15 minutes, equivalent to 30 seconds per question), and a prohibition on using additional information sources. Responses were recorded in an electronic testing system, which automatically registered the chosen option and response time. The resulting data were exported to an Excel spreadsheet for subsequent analysis.

On November 4, 2025, responses from the large language models to the test tasks were collected via the internet. For this purpose, a single prompt with the following content was sequentially submitted to each model through its standard web-based user interface (chat): “You are an anesthesiologist and intensive care specialist. You need to pass a test in your specialty. The test includes questions on the pathophysiology, diagnosis, and treatment of urgent conditions, legal aspects of emergency medical care, and the interpretation of laboratory tests and instrumental studies. The test consists of 30 questions with a single correct answer. Provide a final answer for each question and present a complete list. In your response, indicate only the letter corresponding to your choice, without any explanation”. Following this prompt, the questions (as listed in Appendix 1) were input into the models. As the testing was conducted through a standard web interface, there was no direct access to adjust the “temperature” parameter; the default model settings were used.

For each LLM, ten identical and independent queries were performed to assess the reproducibility and stability of the results. This approach minimizes the impact of random variations in model outputs arising from the stochastic nature of text generation, providing a more representative assessment of their actual capabilities. All responses received from the ten repetitions for each model were com-

pared into an Excel spreadsheet for subsequent comparative analysis.

### Evaluation of results

To objectively compare the effectiveness of the large language models and the actual participants of the “Professionals” competition in completing the test tasks, a multi-level analytical assessment was conducted, incorporating several complementary statistical approaches.

#### 1. Assessment of LLM Position Among Participating Teams (Rank Analysis).

The number of correct answers for each of the 38 teams was calculated. After sorting the results in descending order, the scores of each language model were added to this list, treating the model as a “surrogate participant”. We determined:

- LLM Rank: The position the model would occupy among the teams.
- LLM Percentile: The proportion of teams whose results were lower than the model’s result.

The percentile was calculated using the formula: percentile = (number of teams scoring fewer points than the LLM)/38.

#### 2. Comparison of the LLM Result with the Median Team Result and the Top Team’s Result.

The median score was calculated for all teams. Subsequently, the score of each model was compared to this median score and to the score of the competition’s leading team.

#### 3. Pairwise Comparison of LLM and Team Results (Win-Rate).

Additionally, the pairwise win-rate was calculated, reflecting the probability that an LLM would provide a more correct answer than a randomly selected team. For this, the LLM’s answers were compared with those of every team on a question-by-question basis:

- If the model answered correctly and the team answered incorrectly, it was recorded as a “win” for the model.
- If the model answered incorrectly and the team answered correctly, it was recorded as a “loss”.
- Matching answers (both correct or both incorrect) were considered neutral outcomes and were not included in the calculation.

For each “LLM–team” pair, an individual win-rate was calculated using the formula:

$$\text{Win-rate} = \frac{\text{number of “wins”}}{\text{number of “wins”} + \text{number of “losses”}}$$

The final win-rate for an LLM was determined as the arithmetic mean of these individual win-rates across all teams. In cases where a specific “LLM–team” pair had no wins or losses (a complete tie), that pair was excluded from the calculation of the final indicator.

#### 4. Analysis of Response Agreement Between Models.

To assess differences in the response patterns between large language models, we calculated the observed agree-

ment, which reflects the proportion of identical answers between two models. Additionally, Cohen's  $\kappa$  coefficient, a measure of agreement that corrects for chance agreement, was determined. The interpretation of  $\kappa$  followed the Landis & Koch (1977) scale, where a  $\kappa > 0.81$  is considered "almost perfect agreement".

The nonparametric Kruskal-Wallis test was used to assess differences in the final scores among the LLMs.

#### 5. Correlation Analysis of LLM Responses with Team Responses.

For each question, the "majority answer" was determined — the option selected by the largest number of teams. For all LLMs, the  $\phi$ -coefficient (phi-coefficient) was calculated between the binary vector of the model's answers (correct/incorrect relative to the majority choice) and the binary vector of the majority's answers (consistently correct). In situations where there was no variability (i.e., all values in one of the vectors were identical), the  $\phi$ -coefficient was not calculated, and the observed agreement was used instead.

Statistical data processing was performed using Microsoft Excel 2019 and Python 3.10. As the analyzed indicators were discrete and categorical (number of correct answers, ranks, proportions of matches, agreement coefficients), the following descriptive statistics were used: median, percentiles, rank characteristics, and proportions of matching answers. The level of statistical significance for tests involving the calculation of a  $p$ -value was set at  $p < 0.05$ .

## Results

### General characteristics of the participating teams results

The analysis included the results of the online qualifying test from 38 teams. The number of correct answers on a 30-point scale ranged from 12 to 30, with a median score of 24.5 (25th percentile: 21 points; 75th percentile: 27 points). Only one team out of the 38 achieved the maximum possible result of 30 correct answers.

The results of six large language models were compared with the distribution of test scores from the participating teams. Four models — GPT-4o, GPT-5, Gemini 2.5 Flash, and DeepSeek V-3.2 — demonstrated 100 % accuracy, matching the maximum score of the leading team (30 correct answers). Upon repeated queries, these models produced identical responses in all instances. Based on ten repeated requests, the Qwen3-Max model achieved an average score of 29.9, providing 29 correct answers (making one error on question No. 11 in a single query) and 30 correct answers in the other nine queries. The Alisa AI model provided 29 correct answers across all ten queries, consistently making the same error on question No. 11. Comparative analysis revealed that all six LLMs significant-

ly surpassed the median performance of the participating teams (24.5 correct answers), demonstrating an advantage ranging from +4.5 to +5.5 points. GigaChat failed to provide answers to the test questions in all instances, which should be considered an important finding reflecting the developer's safety policy. In all ten queries, instead of providing answers to the test questions, the system consistently returned a standard safety restriction message: "Generative language models do not have their own opinion — their responses are a generalization of information that is publicly available. To avoid mistakes and misinterpretation, conversations on sensitive topics may be restricted". Summary data are presented in Table 1.

To assess the relative standing of the models among the actual competition participants, each LLM was integrated into the overall ranked list alongside the 38 teams of resident physicians. Five large language models — GPT-4o, GPT-5, DeepSeek V-3.2, Gemini 2.5 Flash, and Qwen3-Max — were assigned the 1st rank, achieving a percentile of 97 %. This signifies superiority over 37 of the 38 teams. In the rank analysis, the Alisa AI model secured 2nd place with a percentile of 92 %, corresponding to an advantage over 35 of the 38 teams. Thus, Alisa AI demonstrated high accuracy, substantially exceeding the median team performance, although it was surpassed by the models that completed the test without errors.

Pairwise comparison of the results revealed a pronounced advantage of the LLMs over the resident teams. For all models exhibiting high accuracy (GPT-4o, GPT-5, DeepSeek V-3.2, Gemini 2.5 Flash, and Qwen3-Max), the win-rate was near-maximal: in most pairwise comparisons, the models provided the correct answer in instances where teams erred. The win-rate for these models was 1.00, indicating a 100 % probability of the model outperforming a randomly selected team. The Alisa AI model had a slightly lower win-rate of 0.97. Nevertheless, even the model with the lowest number of correct answers demonstrated a clear advantage over the majority of participating teams.

Analysis of inter-model agreement showed that four models — GPT-4o, GPT-5, DeepSeek V-3.2, and Gemini 2.5 Flash — demonstrated perfect response agreement (observed agreement = 1.00). Cohen's  $\kappa$  coefficient for all pairwise comparisons between these models was also 1.00, which, according to the Landis & Koch (1977) scale, is interpreted as "almost perfect agreement". The Qwen3-Max and Alisa AI models differed on only one question. The observed agreement between them was 0.97, and Cohen's  $\kappa$  was 0.93, also indicating "almost perfect agreement".

The nonparametric Kruskal-Wallis test was employed to evaluate potential statistical differences among the LLMs' final scores. The four leading models consistently achieved the maximum score, Qwen3-Max scored between 29 and 30 correct answers (average 29.9), and Alisa AI consistently scored 29 correct answers. Despite the presence of inaccuracies in the responses of the latter two models, **the Kruskal-Wallis test revealed no statistically signif-**

icant differences among any of the models ( $p > 0.05$ ). This result, attributable to the extremely narrow range of scores and the limited sample size (6 LLMs, of which 4 had identical results), indicates a similarly high level of effectiveness among the studied models.

To evaluate the alignment between LLM responses and those of the competition participants, a correlation analysis was performed using the  $\phi$ -coefficient. The analysis demonstrated an extremely high degree of concordance for the GPT-4o, GPT-5, DeepSeek V-3.2, Gemini 2.5 Flash, and Qwen3-Max models ( $\phi = 1.00$ ). This reflects perfect agreement between their answers and the majority choice of the teams across all 30 questions. The Alisa AI model yielded a  $\phi$ -coefficient of 0.96, which also corresponds to a very high level of concordance.

A summary of the analysis results is presented in Table 2.

## Discussion

Our study represents one of the first works in the Russian language to compare the performance on specialized medical tests between large language models and teams of resident physicians. Our results demonstrate that modern LLMs are capable of solving single-answer multiple-choice questions in anesthesiology and intensive care in Russian with high accuracy. Four models — GPT-4o, GPT-5, Gemini 2.5 Flash, and DeepSeek V-3.2 — achieved 100 % correct answers, exceeding the median score of the resident teams and matching the maximum score attained by only one team. These findings are consistent with international studies in which LLMs have demonstrated strong performance on test-based assessments. In particular, the study by Kung et al. reported that the GPT-3.5 model achieved a score at or near the passing threshold on

**Table 1.** Results of the test tasks completed by the participants of the "Professionals" competition and large language models

Participant / LLM Model	Type	Total score	Rank among teams	Percentage of correct answers (%)	Difference from Team Median (24.5 points)	Difference from Top Team (30 points)
All teams ( $n = 38$ )	Contest participants	participants	—	12-30-40-100 (81.7 %)*	0	—
Lead team	Contest participants	30	1	100	+5.5	—
GPT-4o	LLM	30	1	100	+5.5	0
GPT-5	LLM	30	1	100	+5.5	0
Gemini 2.5 Flash	LLM	30	1	100	+5.5	0
DeepSeek V-3.2	LLM	30	1	100	+5.5	0
Qwen3-Max	LLM	29.9**	1	99.67 %	+5.4	-0.1
Alisa AI	LLM	29	2	96.67 %	+4.5	-1
GigaChat	LLM	0***	-0	0	—	—

**Note:** \* — median (Me) percentage of correct answers; \*\* — total score calculated as the average of 10 identical queries; \*\*\* — responses not provided by the model.

**Table 2.** Metrics of large language models

Model	Final score	Percentile (%)	Win-rate	Observed agreement	Cohen's $\kappa$	$\phi$ -coefficient
GPT-4o	30	97	1.00	1.00	1.00	1.00
GPT-5	30	97	1.00	1.00	1.00	1.00
DeepSeek V-3.2	30	97	1.00	1.00	1.00	1.00
Gemini 2.5 Flash	30	97	1.00	1.00	1.00	1.00
Qwen3-Max	29.9*	97	1.00	0.97	0.93	1.00
Alisa AI	29	92	0.97	0.97	0.93	0.96
GigaChat	0**	—	—	—	—	—

**Note:** \* — average score based on 10 identical queries; in 1 out of 10 requests, the model provided an incorrect answer to question No. 11; \*\* — responses were not provided by the model.

USMLE-style questions [10]. A subsequent investigation by Gilson et al. using GPT-4 further demonstrated that large language models can outperform the average scores of examinee physicians [8].

Our study provides unique data from a direct, head-to-head comparison of well-known, contemporary LLMs (GPT-4o, GPT-5, Alisa AI, DeepSeek V-3.2, GigaChat, Gemini 2.5 Flash, Qwen3-Max) under uniform conditions within the Russian Federation. The Cohen's  $\kappa$  coefficient of  $\approx 1.00$  among the leading models suggests that their training incorporated up-to-date and consistent medical information. In contrast, models such as Alisa AI (Yandex) and Qwen3-Max (Alibaba) yielded marginally lower, yet still high, scores (96.67 %), with Alisa AI demonstrating a stable, recurring error. This points to potential localized gaps or biases within their respective training datasets. A comparable degree of response consistency between LLMs was previously noted by Mishra et al. [16], who, in their analysis of responses from ChatGPT, Gemini, and Copilot, reported Cohen's  $\kappa$  values ranging from 0.47 to 0.84. Similarly, He Z. et al. [14] found that different LLMs exhibited high similarity in their responses when generating medical advice, although the observed inter-model agreement was lower than that reported in our current analysis.

Application of the Kruskal-Wallis test revealed no statistically significant differences among the models ( $p > 0.05$ ). This finding aligns with the results of Singhal K. et al. [7], who also reported an absence of significant performance variation between models when accuracy spreads were minimal. Thus, our data corroborate that modern LLMs achieve not only high but also remarkably stable levels of concordance, often surpassing those documented in earlier studies. Concurrently, GigaChat's refusal to provide answers underscores the challenge posed by ethical and regulatory constraints, which can artificially limit the utility of LLMs for accessing medical knowledge — a point also raised in the work by Hadweh P. et al. [6].

The high pairwise win-rate observed for all models (ranging from 0.97 to 1.00) further substantiates their advantage over the majority of human participants. This result is consistent with the findings of He Z. et al. [14], who established similar win-rates for the GPT-3.5 and GPT-4 models when their responses were compared against those of other LLMs and users on a web forum in the context of interpreting laboratory results. A systematic review by Liu M. et al. [15] further supports this, demonstrating that GPT-4 achieves an average of 81 % correct answers on medical licensing exam questions and, in most studies, surpasses the average performance of medical students.

To assess the alignment between LLM responses and those of the competition participants, a correlation analysis was performed. The high  $\phi$ -coefficient values obtained in this study ( $\phi = 1.00$  for GPT-4o, GPT-5, DeepSeek V-3.2, Gemini 2.5 Flash, and Qwen3-Max;  $\phi = 0.96$  for Alisa AI) indicate an almost perfect concordance between the models' response patterns and the collective choices of the res-

ident physicians. This strong alignment may, in part, be attributable to the testing format, which consisted of standardized questions with a single unequivocally correct answer. Our data are fully consistent with the results of previously published work [7, 14, 16]. For instance, He Z. et al. [14] also found that GPT-4 and GPT-3.5 responses showed substantial agreement with the majority choices of medical students, although the absolute concordance values were lower than those observed in our study. Mishra V. et al. [16] similarly reported that GPT-4 demonstrated high concordance with physician selections, but noted greater inter-task variability than in our analysis, a finding they attributed to task heterogeneity and varying levels of difficulty. The results of our correlation analysis confirm a high degree of concordance between LLM outputs and the collective choices of the resident teams. This reinforces the clinical relevance of the models and their capacity to approximate the most likely consensus answer of specialists-in-training in anesthesiology and intensive care when addressing single-answer multiple-choice questions.

### Study limitations

Despite the compelling results, this study has several limitations that warrant consideration when interpreting its conclusions.

1. **Limited Scope and Format of Tasks.** The investigation was based on a 30-question test with a single correct answer per item. This format does not permit a comprehensive assessment of an LLM's capacity for complex clinical reasoning, such as analyzing detailed case vignettes, weighing multiple probable diagnoses, or deliberating over various treatment strategies — skills that are integral to real-world clinical practice.
2. **Lack of Analysis of Response Justification.** The study methodology required LLMs to provide only the letter corresponding to their chosen answer, without any accompanying explanation. Consequently, it is impossible to determine whether a correct answer resulted from a genuine understanding of the underlying pathophysiology or from the rote reproduction of memorized information. In both educational and clinical settings, the rationale supporting a decision is often as important as the decision itself.
3. **Topicality and Reliability of Information.** A significant factor limiting the applicability of large language models in medicine is the challenge of ensuring the relevance and reliability of the information they provide. While some contemporary LLMs are architecturally capable of accessing real-time data from the internet, this feature is neither universally available nor guaranteed to be active for every query. By default, most models operate using a static knowledge base, reflective of the data available at the time of their last training update. However, even when web search functionality is technically accessible, a key limitation persists: LLMs

lack intrinsic mechanisms for expert clinical validation. They do not critically appraise the sources of information for concordance with current national and international clinical guidelines, creating a risk of generating responses based on outdated, unreliable, or contradictory data.

4. **Potential Data Leakage.** There is a possibility that test items similar to those used in this study were present in the datasets on which the LLMs were trained. Such data leakage could artificially inflate the models' performance relative to the residents, who had not previously seen the exact questions.

## Conclusion

This study demonstrated that modern large language models achieve high accuracy in solving standardized test

items in anesthesiology and intensive care in Russian, with their results substantially exceeding the median performance of resident physicians. Four models (GPT-4o, GPT-5, DeepSeek V-3.2, Gemini 2.5 Flash) attained the maximum possible score, while the remaining models (Qwen3-Max, Alisa AI) exhibited only isolated errors, maintaining near-perfect agreement with the collective choices of the specialist teams. Analyses of inter-model agreement, pairwise comparisons, and correlations confirmed the high stability and uniformity of LLM responses. However, it is important to note that this level of accuracy does not guarantee that all responses are necessarily derived from the most current and relevant clinical guidelines or data. These findings align with the results of international studies and underscore the potential of LLMs as decision-support tools in medical education.

**Disclosure.** The authors declare no competing interests.

**Author contribution.** All authors according to the ICMJE criteria participated in the development of the concept of the article, obtaining and analyzing factual data, writing and editing the text of the article, checking and approving the text of the article.

**Ethics approval.** The study design was not reviewed by an ethics committee, as the work is based on the analysis of open data comparing the results of large language

models with the responses of participants in an educational competition. This analysis did not involve the use of personal data or any interference with the process of medical care.

**Funding source.** This study was not supported by any external sources of funding.

**Data Availability Statement.** The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Author's ORCID:

Klimov A. A. — 0000-0001-7633-6852

Karelin A.V. — 0009-0000-6667-6407

Liapustin S.V. — 0009-0001-8566-1494

Rudnitsky S.I. — 0000-0001-7458-7893

Tolstova M.A. — 0009-0003-9274-2836

Shamonin A.E. — 0009-0004-7342-0871

Subbotin V.V. — 0000-0002-0921-7199

## References

- [1] Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need. *Adv Neural Inf Process Syst.* 2017; 30: 5998–6008. DOI: 10.48550/arXiv.1706.03762
- [2] Radford A., Narasimhan K., Salimans T., et al. Improving Language Understanding by Generative Pre-Training. OpenAI Technical Report. 2018. DOI: 10.48550/arXiv.1801.06146
- [3] Jiang F., Jiang Y., Zhi H., et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017; 2(4): 230–43. DOI: 10.1136/svn-2017-000101
- [4] Patel B.N., Rosenberg L., Willcox G., et al. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digit Med.* 2019; 2: 111. DOI: 10.1038/s41746-019-0189-7
- [5] Thirunavukarasu A.J., Ting D.S.J., Elangovan K., et al. Large language models in medicine. *Nat Med.* 2023; 29(8): 1930–40. DOI: 10.1038/s41591-023-02448-8
- [6] Hadweh P., Niset A., Salvagno M., et al. Machine Learning and Artificial Intelligence in Intensive Care Medicine: Critical Recalibrations from Rule-Based Systems to Frontier Models. *J Clin Med.* 2025; 14(12): 4026. DOI: 10.3390/jcm14124026

- [7] Singhal K., Tu T., Gottweis J., et al. Toward expert-level medical question answering with large language models. *Nat Med.* 2025; 31(3): 943–50. DOI: 10.1038/s41591-024-03423-7
- [8] Gilson A., Safranek C.W., Huang T., et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ.* 2023; 9: e45312. DOI: 10.2196/45312
- [9] Artsi Y., Sorin V., Konen E., et al. Large language models for generating medical examinations: systematic review. *BMC Med Educ.* 2024; 24(1): 354. DOI: 10.1186/s12909-024-05239-y
- [10] Kung T.H., Cheatham M., Medenilla A., et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health.* 2023; 2(2): e0000198. DOI: 10.1371/journal.pdig.0000198
- [11] American Board of Anesthesiology. Exam Blueprints. Raleigh, NC: The ABA; 2024.
- [12] Brogly N., Engelhardt W., Hill S., et al. European Diploma in Anaesthesiology and Intensive Care in Spain: Results for the part 1 and part 2 exams in the last five years. Are we going in the right direction? *Diploma Europeo en Anestesiología y Cuidados Intensivos en España: resultados de los exámenes de las partes 1 y 2 de los últimos cinco años. ¿Vamos por el buen camino?* *Rev Esp Anestesiol Reanim (Engl Ed).* 2019; 66(4): 206–12. DOI: 10.1016/j.redar.2018.12.009
- [13] Федеральный методический центр аккредитации. Анестезиология-реаниматология: оценочные средства. Методический центр аккредитации специалистов. Москва; 2025. Доступно по ссылке: [https://fmza.ru/fos\\_primary\\_specialized/Anesteziologiya-reanimatologiya/](https://fmza.ru/fos_primary_specialized/Anesteziologiya-reanimatologiya/) (дата обращения: 5.11.2025 г.)
- [14] He Z., Bhasuran B., Jin Q., et al. Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study. *J Med Internet Res.* 2024; 26: e56655. DOI: 10.2196/56655
- [15] Liu M., Okuhara T., Chang X., et al. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. *J Med Internet Res.* 2024; 26: e60807. DOI: 10.2196/60807
- [16] Mishra V., Lurie Y., Mark S. Accuracy of LLMs in medical education: evidence from a concordance test with medical teacher. *BMC Med Educ.* 2025; 25(1): 443. DOI: 10.1186/s12909-025-07009