








## ВОПРОСЫ ОБРАЗОВАНИЯ В АНЕСТЕЗИОЛОГИИ И РЕАНИМАТОЛОГИИ








## ISSUES OF EDUCATION IN ANESTHESIOLOGY AND CRITICAL CARE MEDICINE

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

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

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

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. Шамолин <sup>1</sup>,  
В.В. Субботин <sup>1,3</sup>

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>

<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

<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 Rehabilitation, Moscow, Russia

#### Реферат

#### Abstract

**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 к Коэна) и корреляцион-

**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



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%-ю точность (30 баллов) при подготовке ответов, разделив 1-е ранговое место с командой-лидером. Перцентиль 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

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

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

## Введение

В последние годы широкое распространение в мире получили большие языковые модели (large language models — LLM), ставшие важным этапом развития циф-

ровых технологий. LLM представляют собой системы искусственного интеллекта (ИИ), обученные на больших объемах текстовой информации и способные понимать смысл и контекст человеческой речи. Основой таких систем является архитектура трансформеров (transformer

architecture), предложенная в 2017 г. [1], которая, благодаря механизму внимания (attention mechanism), позволяет анализировать текст целиком и формировать логичные последовательные ответы. В 2018 г. на основе этой технологии была разработана первая модель серии GPT (Generative Pre-trained Transformer) [2].

Публикации, посвященные возможностям применения LLM в здравоохранении, появились в начале 2020-х гг. Уже тогда исследователи стали отмечать перспективность использования данного формата ИИ в медицине [3, 4]. В наши дни модели LLM уже используют для анализа медицинской документации, автоматической интерпретации лабораторных и инструментальных данных с формированием структурированного отчета, а также поддержки принятия клинических решений. Кроме того, модели ИИ демонстрируют высокую точность при анализе больших массивов медицинских данных, включая электронные истории болезни и результаты лабораторных исследований, что открывает перспективы для их использования в клинических информационных системах [5, 6]. Особый интерес представляет оценка возможностей ИИ в рамках специальности «Анестезиология и реаниматология», где ключевое значение имеют скорость анализа информации, принятие решений в условиях ограниченного времени и слаженное командное взаимодействие.

Отдельным направлением применения LLM является медицинское образование и подготовка специалистов [7]. Модели ChatGPT и их аналоги используют в качестве интерактивного инструмента для объяснения сложных понятий, разбора клинических сценариев и тренировки навыков клинического мышления. Применение таких систем способствует повышению эффективности обучения, позволяет студентам и ординаторам самостоятельно анализировать ошибки и получать мгновенные пояснения к вопросам [7, 8]. LLM также применяют для симуляции экзаменационных сценариев и проверки знаний по основным дисциплинам медицины, включая анатомию, фармакологию и патофизиологию [8, 9].

Для объективной оценки уровня знаний и клинического мышления в данной области как в России, так и за рубежом традиционно применяются тестовые задания с одним правильным ответом, а также клинические задачи, моделирующие реальные ситуации оказания медицинской помощи [8–10]. Подобные форматы используются в экзаменах Американского совета по анестезиологии (American Board of Anesthesiology — ABA), Европейского общества анестезиологии и интенсивной терапии (European Society of Anaesthesiology and Intensive Care — ESAIC) в рамках получения Европейского диплома по анестезиологии и интенсивной терапии (European Diploma in Anaesthesiology and Intensive Care — EDAIC), а также в системе аккредитации и итоговой аттестации врачей в Российской Федерации [11–13]. В работах, посвященных анализу

результатов экзаменов USMLE (United States Medical Licensing Examination) и аналогичных сертификационных тестов, показано, что результаты современных больших языковых моделей сопоставимы с ответами врачей-ординаторов и врачей со стажем, а в отдельных случаях даже превосходят их [8, 10].

В октябре 2025 г. в рамках ежегодного Форума анестезиологов-реаниматологов России (ФАРР-2025) в третий раз прошел конкурс ординаторов «Профессионалы» по специальности «Анестезиология и реаниматология». В нем приняли участие 38 команд из разных регионов Российской Федерации, состоящих из врачей-ординаторов первого и второго года обучения. Конкурс включал решение тестовых заданий на отборочном этапе, а также выполнение заданий по специальности и клинических ситуационных задач, максимально приближенных к реальной практике в формате очной викторины.

После проведения конкурса у организаторов возник интерес сопоставить результаты участников с ответами, которые могут дать современные большие языковые модели (GPT-4o, GPT-5, Alisa AI, DeepSeek V-3.2, GigaChat, Gemini 2.5 Flash). Эта идея появилась вследствие целого ряда наблюдений, продемонстрировавших, что многие LLM уже показывают высокий уровень успешности при решении тестовых заданий по медицине. Однако остается открытым вопрос: способны ли такие модели адекватно отвечать на узкоспециализированные вопросы. В то же время эффективность применения больших языковых моделей при решении медицинских заданий на русском языке остается практически неизученной. Большинство опубликованных исследований выполнено на английском языке, что не позволяет напрямую экстраполировать их результаты на условия клинической практики в Российской Федерации.

## Цель исследования

Оценить успешность выполнения тестовых заданий с одним вариантом правильного ответа на русском языке по специальности «Анестезиология и реаниматология» современными большими языковыми моделями (GPT-4o, GPT-5, Alisa AI, DeepSeek V-3.2, GigaChat, Gemini 2.5 Flash, Qwen 3-Max) в сравнении с результатами команд врачей-ординаторов, участвовавших в конкурсе «Профессионалы», прошедшем в рамках Форума анестезиологов-реаниматологов России 2025 г.

## Материалы и методы

### Дизайн исследования

Проведено сравнительное исследование ответов на тестовые задания отборочного этапа кон-

курса «Профессионалы» 2025 г. по специальности «Анестезиология и реаниматология». Оценивали результаты реальных участников конкурса и современных больших языковых моделей. Исследование включало анализ ответов на 30 тестовых заданий с одним правильным ответом, охватывающих патофизиологию, диагностику и лечение неотложных состояний, правовые аспекты экстренной помощи, интерпретацию лабораторных и инструментальных данных.

Анализ дизайна исследования в этическом комитете не проводили, так как работа основана на анализе открытых данных, сравнении результатов больших языковых моделей с ответами участников образовательного конкурса, не предусматривающем использование персональных данных или вмешательство в процесс оказания медицинской помощи.

### Тестовые материалы

В данном исследовании использовали тест отборочного этапа конкурса «Профессионалы», состоящий из 30 вопросов, который был разработан методическим комитетом конкурса в соответствии с требованиями к теоретической подготовке врачей-ординаторов по специальности «Анестезиология и реаниматология».

### Участники исследования

Реальными участниками исследования являлись врачи-ординаторы по специальности «Анестезиология и реаниматология» первого или второго года обучения, прошедшие регистрацию для участия в конкурсе. В анализ включены результаты тестирования 38 команд, каждая из которых состояла из четырех человек.

В исследовании использованы ответы на тесты следующих больших языковых моделей: GPT-4o (IT-компания OpenAI), GPT-5 (IT-компания OpenAI), Alisa AI (IT-компания Yandex), DeepSeeK V-3.2 (IT-компания DeepSeek), GigaChat (компания Сбер), Gemini 2.5 Flash (компания Google), Qwen3-Max (IT-компания Alibaba).

### Методы

15 сентября 2025 г. для участников конкурса «Профессионалы» было проведено онлайн-тестирование на платформе МТС LINK. Тестирование было организовано в условиях, максимально приближенных к экзаменационным: онлайн-формат, ограничение времени (15 мин, что соответствовало 30 с на вопрос) и запрет на использование дополнительных источников информации. Ответы фиксировались в электронной системе тестирования с автоматической регистрацией выбранного варианта и времени ответа. Полученные данные были экспортированы в формат таблицы Excel для последующего анализа.

4 ноября 2025 г. был выполнен сбор ответов больших языковых моделей на тестовые задания через сеть Интернет. Для этого каждой модели через стандартный веб-интерфейс пользователя (чат) последовательно задавали единый промпт следующего содержания: «Вы врач — анестезиолог-реаниматолог. Вам нужно пройти тест по специальности. Тестирование включает вопросы по патофизиологии, диагностике и лечению urgentных состояний, правовым аспектам оказания экстренной медицинской помощи, интерпретации данных лабораторных анализов и инструментальных исследований. Тест состоит из 30 вопросов с одним правильным ответом. Дайте окончательный ответ на каждый вопрос и представьте готовый список. В ответе укажите только букву, без объяснений». После этого в модели загружали вопросы из приложения 1. Поскольку тестирование проводили через стандартный веб-интерфейс, прямой доступ к настройке параметра temperature отсутствовал. Использовались настройки модели по умолчанию.

Для каждой LLM было выполнено по 10 идентичных независимых запросов с целью оценки воспроизводимости и стабильности результатов. Такой подход позволяет минимизировать влияние случайных вариаций в работе моделей, возникающих вследствие стохастической природы генерации текста, и получить более репрезентативную оценку их фактических возможностей. Все полученные ответы в количестве 10 повторов для каждой модели были перенесены в таблицу Excel для последующего сравнительного анализа.

### Оценка результатов

Для объективного сравнения эффективности выполнения тестовых заданий большими языковыми моделями и реальными участниками конкурса «Профессионалы» проведена многоуровневая аналитическая оценка, включающая несколько взаимодополняющих статистических подходов:

1. Оценка положения LLM среди команд участников (ранговый анализ).

Подсчитывали количество правильных ответов каждой из 38 команд. После сортировки результатов в порядке убывания в этот список добавляли показатели каждой языковой модели, рассматривая ее как «условного участника». Определяли:

- ранг LLM — место, которое модель занимала среди команд;
- процентиль LLM — долю команд, чьи результаты ниже результата модели.

Процентиль рассчитывали по формуле:

$$\text{Процентиль} = (\text{число команд, набравших меньше баллов, чем LLM}) / 38.$$

2. Сравнение результата LLM с медианным результатом всех команд и результатом команды-лидера.

Для всех команд вычисляли медианный результат. Далее сравнивали показатели каждой модели с медианой.

ной суммы баллов и результатом команды-лидера конкурса.

### 3. Парное сравнение результатов LLM и команд.

Дополнительно рассчитывали долю «побед» при парном сравнении (pairwise win-rate), отражающую вероятность того, что LLM даст более корректный ответ, чем случайно выбранная команда. Для этого проводили сравнение ответов LLM с ответами всех команд:

- если модель давала правильный ответ, а команда — неправильный, фиксировали «победу» модели;
- если модель ошибалась, а команда отвечала правильно — «поражение»;
- совпадение ответов (оба правильные либо оба неправильные) расценивали как нейтральный исход и не включали в расчет.

Для каждой пары «LLM – команда» вычисляли индивидуальный показатель доли «побед» по формуле:

Доля «победы» = число «побед» / (число «побед» + число «поражений»).

Итоговый показатель для LLM определяли как среднее арифметическое этих индивидуальных долей по всем командам. В случаях, когда для конкретной пары «LLM – команда» не было ни «побед», ни «поражений» (полная ничья), такая пара исключалась из расчета итогового показателя.

### 4. Анализ согласованности ответов между моделями.

Для оценки различий в структуре ответов между большими языковыми моделями рассчитывали показатель наблюдаемой согласованности (observed agreement), отражающий долю совпадающих ответов между двумя моделями. Дополнительно определяли коэффициент к Коэна, показатель согласованности с поправкой на случайные совпадения. Интерпретацию к выполняли согласно шкале Landis & Koch (1977):  $k > 0,81$  расценивали как «почти полное согласие».

Для оценки различий между итоговыми показателями LLM применяли непараметрический критерий Краскела–Уоллиса.

### 5. Корреляционный анализ ответов LLM с ответами команд.

Для каждого вопроса определяли ответ большинства вариант, выбранный наибольшим числом команд. Для всех LLM рассчитывали коэффициент  $\phi$  (phi-coefficient) между бинарным вектором ответов модели и бинарным вектором ответов большинства. В ситуациях, когда отсутствовала вариабельность (все значения в одном из векторов совпадали), коэффициент  $\phi$  не рассчитывали и использовали показатель наблюдаемой согласованности.

Статистическую обработку данных выполняли с использованием Microsoft Excel 2019 и Python 3.10. Поскольку анализируемые показатели являлись дискретными и категориальными (число правильных ответов, ранги, доли совпадений, коэффициенты согласованности), применяли следующие показатели описательной статистики: медиану, перцентили, ранговые

характеристики, долю совпадающих ответов. Уровень статистической значимости для тестов, предполагающих вычисление  $p$ -значения, принимали равным  $< 0,05$ .

## Результаты

### Общая характеристика результатов конкурсных команд

В анализ включены результаты отборочного онлайн-тестирования 38 команд. Количество правильных ответов по 30-балльной шкале варьировало от 12 до 30 баллов, медианный результат составил 24,5 балла (25-й перцентиль — 21 балл, 75-й перцентиль — 27 баллов). Только одна команда из 38 показала максимальный возможный результат (30 правильных ответов).

Результаты шести больших языковых моделей были сопоставлены с распределением тестовых баллов команд-участников. Четыре модели GPT-4o, GPT-5, Gemini 2.5 Flash и DeepSeek V-3.2 продемонстрировали 100 % правильных ответов, полностью повторив максимальный результат команды-лидера (30 правильных ответов). При повторных запросах модели демонстрировали идентичные ответы во всех случаях. Модель Qwen3-Max, по данным 10 повторных запросов, обеспечила усредненный балл 29,9, предоставив 29 правильных ответов (допустив одну ошибку в 11-м вопросе в одном запросе) и 30 правильных ответов в девяти запросах. Модель Alisa AI во всех десяти запросах обеспечила 29 правильных ответов, каждый раз допуская одну и ту же ошибку в вопросе № 11. Сравнительный анализ показал, что все шесть LLM существенно превосходили медианный уровень команд-участников (24,5 правильного ответа), демонстрируя преимущество от +4,5 до +5,5 балла. Модель GigaChat не смогла предоставить ответы на тест из-за особенностей настройки, что следует рассматривать как важный результат, отражающий политику безопасности разработчика. Во всех десяти запросах вместо ответов на тестовые задания система выдавала стандартное сообщение ограничения: «Генеративные языковые модели не обладают собственным мнением, их ответы являются обобщением информации, находящейся в открытом доступе. Чтобы избежать ошибок и неправильного толкования, разговоры на чувствительные темы могут быть ограничены». Сводные данные представлены в табл. 1.

Для оценки относительной позиции моделей среди реальных участников конкурса каждая LLM была включена в общий ранжированный список наряду с 38 командами врачей-ординаторов. Пяти большим языковым моделям GPT-4o, GPT-5, DeepSeek V-3.2, Gemini 2.5 Flash и Qwen3-MAX было присвоено 1-е ранговое место, а их перцентиль составил 97 %, что соответствует превосходству над 37 из 38 команд. Модель Alisa AI в ранговом анализе заняла 2-е место, а ее перцентиль

**Таблица 1.** Итоги выполнения тестовых заданий участниками конкурса «Профессионалы» и большими языковыми моделями**Table 1.** Results of the test tasks completed by the participants of the "Professionals" competition and large language models

Участник/Модель LLM	Тип	Сумма баллов	Ранг среди команд	Процент правильных ответов (%)	Разница с медианой команд (24,5 балла)	Разница с лучшей командой (30 баллов)
Все команды ( $n = 38$ )	Участники конкурса	12–30	—	40–100 (81,7 %)*	0	—
Команда-лидер	Участники конкурса	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 балл
Alisa AI	LLM	29	2	96,67 %	+4,5	-1 балл
GigaChat	LLM	0***	—	0	—	—

**Примечание:** \* — медиана (Me) процента правильных ответов; \*\* — сумма баллов рассчитана на основании 10 идентичных запросов, \*\*\* — ответы не предоставлены моделью.

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

составил 92 %, что соответствует превосходству над 35 из 38 команд. Модель Alisa AI показала высокую точность, существенно превышающую медианные показатели команд, но уступающую моделям с безошибочным выполнением теста.

Попарное сравнение результатов показало выраженное преимущество больших LLM над командами врачей-ординаторов. Для всех моделей, продемонстрировавших высокий уровень точности (GPT-4o, GPT-5, DeepSeek V-3.2, Gemini 2.5 Flash и Qwen3-MAX), доля «победы» была близка к максимальной: в большинстве парных сравнений модели давали правильный ответ там, где команды ошибались. Показатель доли «победы» у этих моделей составляет 1,00, что соответствует вероятности «победы» моделей в 100 % случаев над случайно выбранной командой. Модель Alisa AI имела более низкий показатель, однако и в этом случае средняя доля «побед» составляет 0,97. Тем не менее модель с наименьшим числом правильных ответов демонстрирует преимущество перед большинством команд участников конкурса.

Анализ показателей наблюдаемой согласованности показал, что четыре модели GPT-4o, GPT-5, DeepSeek V-3.2 и Gemini 2.5 Flash продемонстрировали полное совпадение ответов (observed agreement = 1,00). Коэффициент к Коэна для всех парных сравнений между этими моделями также принимал значение 1,00, что по шкале Landis & Koch (1977) трактуется как «почти полное согласие». Модели Qwen3-MAX и Alisa AI име-

ли расхождение только в одном вопросе. Показатель наблюдаемой согласованности составил 0,97, а коэффициент к Коэна 0,93, что также трактуется как «почти полное согласие».

Для оценки возможных статистических различий между итоговыми результатами LLM был применен непараметрический критерий Краскела—Уоллиса. Четыре модели-лидера показали неизменный максимальный результат, Qwen3-Max — 29–30 правильных ответов (в среднем — 29,9), Alisa AI — стабильные 29 правильных ответов. Несмотря на наличие неточных ответов у двух последних моделей, критерий Краскела—Уоллиса не выявил статистически значимых различий между моделями ( $p > 0,05$ ). Данный результат, обусловленный крайне малым разбросом значений и ограниченным объемом выборки (6 LLM, из которых 4 имеют идентичный результат), свидетельствует о схожей высокой эффективности изученных моделей.

Для оценки совпадения ответов LLM и участников конкурса был проведен корреляционный анализ с использованием коэффициента  $\phi$ . Анализ показал крайне высокую степень согласованности у моделей GPT-4o, GPT-5, DeepSeek V-3.2, Gemini 2.5 Flash и Qwen3-Max ( $\phi = 1,00$ ), что отражает абсолютное совпадение их ответов с выбором большинства команд по всем 30 вопросам. Модель Alisa AI продемонстрировала  $\phi = 0,96$ , что также соответствует очень высокой согласованности.

Сводные данные проведенного анализа представлены в табл. 2.

**Таблица 2.** Показатели больших языковых моделей

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

Модель	Итоговый балл	Процентиль (%)	Доля «победы»	Показатель наблюдаемой согласованности	к Коэна	φ-коэффициент
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**	—	—	—	—	—

**Примечание:** \* — в 1 из 10 повторных запросов система представила неправильный ответ на 11-й вопрос; \*\* — ответы не предоставлены моделью.

**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.

## Обсуждение

Наше исследование представляет собой одну из первых работ на русском языке, в которой было проведено сравнение результатов тестирования по узко-профильной медицинской тематике между большими языковыми моделями и командами врачей-ординаторов. Полученные нами результаты демонстрируют, что современные LLM способны с высокой точностью решать тестовые задания по анестезиологии и реаниматологии с одним правильным ответом на русском языке. Четыре модели GPT-4o, GPT-5, Gemini 2.5 Flash и DeepSeek V-3.2 показали 100 % правильных ответов, превысив медианный результат команд-ординаторов и повторив максимальный балл, достигнутый лишь одной командой. Эти данные согласуются с международными работами, в которых LLM показывали высокие результаты при решении тестовых заданий. В частности, в исследовании T.H. Kung et al. модель GPT-3.5 достигла результата на уровне или около проходного балла при решении тестов USMLE [10]. Более позднее исследование A. Gilson et al. с использованием GPT-4 также продемонстрировало, что большие языковые модели превосходят средние показатели испытуемых врачей [8].

Наше исследование предоставляет уникальные данные по прямому сравнению известных современных LLM на территории РФ (GPT-4o, GPT-5, Alisa AI, DeepSeek V-3.2, GigaChat, Gemini 2.5 Flash, Qwen3-Max) в равных условиях. Показатель к Коэна  $\approx 1.00$  между моделями-лидерами говорит об использовании актуальной медицинской информации при обучении всех вышеперечисленных моделей. При этом такие модели, как Alisa AI (Yandex) и Qwen3-Max (Alibaba), показали результаты лишь незначительно ниже (96.67 %), причем ошибка у Alisa AI была стабильной. Это указывает

на возможные локальные пробелы в обучающих массивах данных. Подобная степень согласованности ответов между LLM ранее отмечалась в работах V. Mishra et al., где при анализе ответов ChatGPT, Gemini и Copilot значения Коэна к варьировали от 0,47 до 0,84 [16]. Сходные данные приводятся в исследовании Z. He et al., где различные модели демонстрировали высокую близость ответов при оценке медицинских советов, однако наблюдаемая схожесть между LLM также была ниже, чем в представленном нами анализе [14].

Применение критерия Краскела—Уоллиса не выявило статистически значимых различий между моделями ( $p > 0,05$ ), что согласуется с результатами исследования K. Singhal et al., где также сообщалось об отсутствии значимых различий между моделями при минимальном разбросе точности [7]. Таким образом, наши данные подтверждают, что современные LLM достигают не только высокого, но и чрезвычайно стабильного уровня согласованности, превосходящего показатели большинства ранее опубликованных исследований. В то же время отказ GigaChat от ответа подчеркивает проблему этических и регуляторных ограничений, которые могут искусственно ограничивать использование LLM для получения медицинских знаний, что также обсуждалось в работе P. Hadweh et al. [6].

Высокий показатель доли «побед» (pairwise win-rate) у всех моделей (0,97–1,00) также подтверждает их преимущество перед большинством реальных участников. Наш результат соотносится с результатами Z. He et al., установивших подобные показатели доли «побед» для моделей GPT-3,5 и GPT-4 при сравнении с другими LLM и ответами пользователей веб-форума при интерпретации лабораторных анализов [14]. Систематический обзор M. Liu et al. также показал, что GPT-4 в среднем достигает 81 % правильных ответов

на вопросах медицинских лицензированных экзаменов и в большинстве работ превосходит средние результаты студентов-медиков [15].

Для оценки совпадения ответов LLM и участников конкурса был проведен корреляционный анализ. Высокие значения коэффициента  $\phi$ , полученные в настоящем исследовании ( $\phi = 1,00$  для GPT-4o, GPT-5, DeepSeek V-3.2, Gemini 2.5 Flash и Qwen3-MAX;  $\phi = 0,96$  для Alisa AI), указывают на практически полное совпадение структуры ответов моделей с коллективным выбором врачей-ординаторов, что, в частности, может быть объяснено форматом тестирования: стандартизированные вопросы с одним правильным вариантом. Наши данные полностью согласуются с результатами ранее опубликованных работ [7, 14, 16]. В исследовании Z. He et al. модели GPT-4 и GPT-3.5 также показали значительное совпадение ответов с большинством студентов-медиков, хотя абсолютные показатели были ниже, чем в нашей работе [14]. V. Mishra et al. также сообщили, что GPT-4 демонстрирует высокое совпадение с вариантами выбора врачей, однако вариабельность между задачами была выше, чем в нашем анализе, что авторы связывали с неоднородностью и разной степенью сложности заданий [16]. Результаты проведенного корреляционного анализа свидетельствуют о практически полном совпадении структуры ответов LLM с ответами большинства команд, что подтверждает высокую клиническую ориентированность моделей и их способность воспроизводить наиболее вероятный коллективный выбор врачей-ординаторов по специальности «Анестезиология и реаниматология» при решении тестовых заданий с одним правильными ответом.

### Ограничения исследования

Несмотря на убедительные результаты, настоящее исследование имеет ряд ограничений, которые необходимо учитывать при интерпретации его выводов.

1. Ограниченный объем и формат заданий. Исследование основано на тесте, состоящем из 30 вопросов с одним верным ответом. Такой формат не позволяет в полной мере оценить способность LLM к комплексному клиническому мышлению, анализу развернутых клинических случаев, взвешиванию нескольких вероятных диагнозов или тактик лечения, что является неотъемлемой частью реальной клинической практики.
2. Отсутствие анализа обоснования ответов. В рамках методологии исследования LLM давали только буквенный вариант ответа без каких-либо объяснений. Следовательно, невозможно оценить, был ли правильный ответ результатом глубокого понимания

патофизиологии или шаблонного воспроизведения информации. В реальной образовательной и клинической практике обоснование решения не менее важно, чем сам ответ.

3. Актуальность и достоверность информации. Существенным фактором, ограничивающим применимость больших языковых моделей в медицине, является проблема актуальности и достоверности информации. Хотя некоторые современные LLM архитектурно способны обращаться к данным в сети Интернет, эта функция не является ни повсеместной, ни гарантированно активированной при каждом запросе. Большинство моделей по умолчанию оперируют статичной базой знаний, сформированной в момент последнего обучения. Однако даже при наличии технической возможности веб-поиска ключевое ограничение сохраняется: LLM лишены встроенных механизмов экспертной клинической валидации. Они не осуществляют критической оценки источников на соответствие текущим национальным и международным рекомендациям, что создает риск генерации ответов на основе устаревших, недостоверных или противоречивых данных.
4. Потенциальная «утечка» данных. Существует вероятность, что похожие тестовые задания могли присутствовать в данных, на которых обучались модели LLM. Это могло искусственно завысить результаты моделей по сравнению с ординаторами.

### Заключение

Проведенное исследование показало, что современные большие языковые модели демонстрируют высокую точность при решении стандартизированных тестовых заданий по анестезиологии и реаниматологии на русском языке, и их результаты значительно превосходят медианные результаты врачей-ординаторов. Четыре модели (GPT-4o, GPT-5, DeepSeek V-3.2, Gemini 2.5 Flash) достигли максимального возможного балла, а остальные (Qwen3-MAX, Alisa AI) показали лишь единичные ошибки, сохраняя почти полное совпадение с коллективным выбором специалистов. Анализ согласованности, попарных сравнений и корреляции подтвердил высокую стабильность и однородность ответов LLM. При этом важно учитывать, что достигнутая точность не гарантирует, что в основе всех ответов лежат самые актуальные и соответствующие современным стандартам данные. Полученные данные согласуются с результатами международных исследований и подчеркивают потенциал LLM как инструмента поддержки решений в медицинском образовании.

**Конфликт интересов.** Авторы заявляют об отсутствии конфликта интересов.

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

**Вклад авторов.** Все авторы в равной степени участвовали в разработке концепции статьи, получении и анализе фактических данных, написании и редактировании текста статьи, проверке и утверждении текста статьи.

**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.

**Этическое утверждение.** Анализ дизайна исследования в этическом комитете не проводили, так как работа основана на анализе открытых данных, сравнении результатов больших языковых моделей с ответами участников образовательного конкурса, не предусматривающем использование персональных данных или вмешательство в процесс оказания медицинской помощи.

**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.

**Информация о финансировании.** Авторы заявляют об отсутствии внешнего финансирования при проведении исследования.

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

**Декларация о наличии данных.** Данные, подтверждающие выводы этого исследования, можно получить у корреспондирующего автора по обоснованному запросу.

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

#### ORCID авторов:

Климов А.А. — 0000-0001-7633-6852

Карелин А.В. — 0009-0000-6667-6407

Ляпустин С.В. — 0009-0001-8566-1494

Рудницкий С.И. — 0000-0001-7458-7893

Толстова М.А. — 0009-0003-9274-2836

Шамонин А.Е. — 0009-0004-7342-0871

Субботин В.В. — 0000-0002-0921-7199

## Литература/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 Anesthesiol 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