Российская оториноларингология. 2025;24(3):53–62
Russian Otorhinolaryngology. 2025;24(3):53-62

ОТИАТРИЯ

Научная статья
УДК 616.284-072.1:519.766.2
https://doi.org/10.18692/1810-4800-2025-3-53-62

Потенциал многомодальной языковой модели для предварительной оценки отоскопических изображений

М. В. Комаров[1], О. И. Гончаров[2], А. А. Федотова[3]

[1] *Санкт-Петербургский научно-исследовательский институт уха, горла, носа и речи, Санкт-Петербург, 190013, Российская Федерация*

[1,3] *Северо-Западный государственный медицинский университет имени И. И. Мечникова, Санкт-Петербург, 195067, Российская Федерация*

[2] *Национальный медицинский исследовательский центр имени В. А. Алмазова, Санкт-Петербург, 197341, Российская Федерация*

[1,2,3] *Городская больница № 26, Санкт-Петербург, 196240, Российская Федерация*

[1] 7_line@mail.ru✉, https://orcid.org/0000-0003-4471-3603
[2] https://orcid.org/0000-0003-3738-4944
[3] https://orcid.org/0000-0002-4816-2098

Реферат. Пилотное исследование оценило возможности универсальной многомодальной LLM ChatGPT o3 для интерпретации отоскопических изображений. В работу включили 38 кадров, разделенных на девять клинических категорий: от нормы и инородных тел до послеоперационных состояний и опухолей среднего уха. Аннотацию «золотого стандарта» обеспечили два эксперта-оториноларинголога (κ > 0,85), при расхождениях — консенсус. Модель обрабатывала каждый кадр в новом сеансе по запросу «Что ты видишь на этой фотографии?» ChatGPT o3 продемонстрировала 100% точности в разделении «норма/патология» (95% CI 90,8–100%), чувствительность и специфичность 100%, PPV/NPV = 100%. Верность формулировки клинического диагноза составила 81,6% (31/38). По пяти ключевым морфологическим признакам (перфорация, экссудат, гиперемия, тимпаносклероз, холестеатома) средний F1-score достиг 0,92, Cohen's κ = 0,87. Экспертная оценка полезности текстовых описаний по 5-балльной шкале показала $M = 4,4 \pm 0,6$, ICC = 0,82; различий между группами не выявлено ($p = 0,24$). Spearman's ρ = 0,72 ($p < 0,001$) подтвердил связь между числом правильно определенных признаков и оценкой полезности. Среднее время отклика 30–40 с. Результаты указывают на высокий потенциал ChatGPT o3 для предварительного скрининга, стандартизации отчетности и обучения. Для клинического внедрения необходимы масштабная проспективная валидация, структурирование вывода и интеграция количественных инструментов.
Ключевые слова: отоскопия, многомодальная языковая модель, ChatGPT o3, диагностика среднего уха, морфологический анализ, скрининг, телемедицина, объяснимый ИИ, точность классификации, межэкспертное согласие

Для цитирования: Комаров М. В., Гончаров О. И., Федотова А. А. Потенциал многомодальной языковой модели для предварительной оценки отоскопических изображений. *Российская оториноларингология.* 2025;24(3):53–62. https://doi.org/10.18692/1810-4800-2025-3-53-62

OTIOLOGY

Science article

Potential of multimodal language model for preliminary evaluation of otoscopic images

M. V. Komarov[1], O. I. Goncharov[2], A. A. Fedotova[3]

[1] *Saint Petersburg Research Institute of Ear, Throat, Nose and Speech, Saint Petersburg, 190013, Russian Federation*

[1,3] *Mechnikov North-Western State Medical University, Saint Petersburg, 195067, Russian Federation*

[2] *Almazov National Medical Research Centre, Saint Petersburg, 197341, Russian Federation*

[1,2,3] *City Hospital No. 26, Saint Petersburg, 196240, Russian Federation*

[1] 7_line@mail.ru✉, https://orcid.org/0000-0003-4471-3603
[2] https://orcid.org/0000-0003-3738-4944
[3] https://orcid.org/0000-0002-4816-2098

Abstract. A pilot study evaluated the capabilities of the universal multimodal LLM ChatGPT 03 for interpreting otoscopic images. Thirty-eight frames were grouped into nine clinical categories—from normal and foreign bodies to postoperative states and middle-ear tumors. A "gold standard" annotation was provided by two otorhinolaryngology experts (Cohen's κ > 0.85), with consensus reached in cases of disagreement. Each frame was processed in a new session with the prompt "What do you see in this photo?" ChatGPT 03 achieved 100% accuracy in distinguishing "normal vs. pathology" (95% CI 90.8–100%), with sensitivity and specificity, PPV/NPV (positive predictive value/negative predictive value) = 100%. The correctness of its clinical diagnosis formulation was 81.6% (31/38). For five key morphological features (perforation, effusion, hyperemia, tympanosclerosis, cholesteatoma), the mean F1-score was 0.92, and Cohen's κ = 0.87. Expert ratings of the utility of its text descriptions on a 5-point scale yielded $M = 4.4 \pm 0.6$, ICC = 0.82, with no significant differences between groups ($p = 0.24$). Spearman's ρ = 0.72 ($p < 0.001$) confirmed a strong positive correlation between the number of correctly identified features and the usefulness assessment. The average response time was 30–40 s. These findings underscore ChatGPT 03's high potential for preliminary screening, report standardization, and education. Clinical implementation will require large-scale prospective validation, structured output, and integration of quantitative tools.

Keywords: otoscopy, multimodal language model, ChatGPT 03, middle ear diagnosis, morphological analysis, screening, telemedicine, explainable AI, classification accuracy, inter-rater agreement

For citation: Komarov M. V., Goncharov O. I., Fedotova A. A. Potential of multimodal language model for preliminary evaluation of otoscopic images. *Russian Otorhinolaryngology*. 2025;24(3):53-62. (In Russ.) https://doi.org/10.18692/1810-4800-2025-3-53-62

Введение

Отоскопия остается фундаментальным инструментом первичной диагностики патологий среднего уха. Ее высокая информативность сочетается с минимальной инвазивностью, однако качество интерпретации отоскопических изображений в значительной мере зависит от опыта и квалификации врача-оториноларинголога [1–3]. В ряде регионов дефицит специалистов приводит к задержкам в диагностике, особенно при сложных случаях (онкологическая настороженность, послеоперационные состояния, экссудативный отит и др.). Автоматизация анализа изображений на базе методов искусственного интеллекта (ИИ) способна повысить доступность и воспроизводимость исследования, снизив нагрузку на клиницистов и ускорив принятие решений в первичном звене [4–7].

До недавнего времени основной упор делался на разработку узкоспециализированных сверточных нейросетей (Convolutional Neural Network — CNN) для классификации отдельных нозологических форм (острый и экссудативный отит, перфорация барабанной перепонки) с достижением чувствительности и специфичности 90–98% [2–4]. Однако такие модели требуют крупные аннотированные датасеты и сложную донастройку (fine tuning) под каждый конкретный сценарий применения [8–12]. Кроме того, они не умеют «объяснять» свои решения на естественном языке, что ограничивает их интеграцию в образовательные и телемедицинские платформы [1, 7–9, 13].

В настоящей работе впервые оценена способность универсальной многомодальной языковой модели ChatGPT 03 (OpenAI), доступной для неограниченного круга пользователей с 16 апреля 2025 года, к клинико-морфологической интерпретации широкого спектра отоскопических изображений. Мы протестировали модель на девяти принципиально разных группах.

Предварительное тестирование показало высокую точность описания ключевых морфологических признаков и своевременное дифференцирование «нормы» и «патологии» в 100% случаев, при этом модель сразу определяла ненормальность отоскопии даже при низкой онкологической настороженности.

Цель исследования

Оценить базовые возможности ChatGPT 03 в анализе отоскопических изображений различных нозологических категорий и определить его потенциал как инструмента предварительного скрининга и дистанционного обучения.

Задачи исследования

Проанализировать полноту и точность описания морфологических изменений в каждом из девяти сценариев.

Сравнить качество интерпретации LLM (Large Language Model) с результатами экспертов-отоларингологов.

Выявить сильные и слабые стороны LLM подхода по критериям клинической полезности и объяснимости.

Полученные данные позволят сформировать рекомендации по интеграции ChatGPT 03 в клинические и образовательные процессы ЛОР-профильных учреждений.

Материалы и методы

Дизайн исследования. Пилотное экспериментальное исследование возможностей языковой модели ChatGPT 03 (OpenAI) для клинико-морфологической интерпретации отоскопических изображений. Исследование выполнено в апреле 2025 года на базах ФГБУ СПб НИИ ЛОР Минздрава России, ФГБУ «НМИЦ им. В. А. Алмазова» Минздрава России, ФГБОУВО «СЗГМУ им. И. И. Мечникова» в оториноларингологическом отделении СПб ГБУЗ «Городская больница № 26».

Формирование и описание выборки. Для пилотного анализа было отобрано 38 эндоскопических отоскопических иллюстраций из клинической базы фотодокументации ЛОР-отделения (ALBUM_CLOUD.pdf, рис. 1). Изображения разделены на девять диагностических групп по нозологии и этапу заболевания/лечения.

– Источник изображений: клиническая база фотодокументации ЛОР-отделения.

– Критерии включения: четкая визуализация pars tensa/pars flaccida или трепанационной полости, отсутствие технических артефактов. Хорошая фокусировка — возможность однозначно оценить морфологические признаки.

– Комментарий по отбору: для каждой группы изображения были выбраны так, чтобы представить диапазон клинических проявлений: от нормы до сложнейших для описания состояний (иллюстрация послеоперационной полости), а также были включены иллюстрации с «ловушками» — мигрировавший электрод, волосы и т. д. Номера иллюстраций соответствуют порядку в приложении ALBUM_CLOUD.pdf, что обеспечивает прозрачность сопоставления золотого стандарта и выходов модели. (рис. 1). Такой сбалансированный (по группам) и жестко отфильтрованный

Таблица 1

Состав и нумерация диагностических групп отоскопических изображений

Table 1

Composition and numbering of diagnostic groups of otoscopic images

Группа	Описание	Количество иллюстраций	№–№ иллюстраций в альбоме
1	Норма (здоровая перепонка)	3	1–3
2	Инородные тела в слуховом проходе	6	4–9
3	Экссудативный отит	6	10–15
4	Посттравматический разрыв барабанной перепонки	4	16–19
5	Острый средний отит	4	20–23
6	Ретракционные карманы барабанной перепонки	5	24–28
7	Опухоли среднего уха	3	29–31
8	Состояния после шунтирования барабанной перепонки	3	32–34
9	Состояния после радикальной операции на ухе (canal wall down)	4	35–38
Всего		38	

Рис. 1. QR-код, содержащий ссылку на скачивание файла в формате *.pdf. Название файла ALBUM_CLOUD.pdf. Файл содержит альбом пронумерованных отоскопических иллюстраций, предлагаемых для анализа LLM. https://disk.yandex.ru/i/j7l7igK5wLU3kw

Fig. 1. QR code containing a link to download the file in *.pdf format. The file is named ALBUM_CLOUD.pdf. The file contains an album of numbered otoscopic illustrations proposed for analysis by an LLM. https://disk.yandex.ru/i/j7l7igK5wLU3kw

набор изображений позволил всесторонне протестировать модель ChatGPT o3 на ключевых диагностических сценариях отоскопии.

Подготовка изображений

– Обрезка и выравнивание фотографии объекта съемки проводились для иллюстраций: 10, 11, 12, 13, 14, 16, 17, 18, 19, 20, 23, 30, 31, 36, 38.

– Предварительная балансировка яркости/контраста, а также очистка фотографии от артефактов не проводились

Эталонная аннотация (золотой стандарт)

– Два независимых эксперта-оториноларинголога (стаж > 10 лет в отохирургии) дали развернутую интерпретацию каждого кадра (наличие/отсутствие ключевых признаков: перфорация, экссудат, ретракция, бляшки, уплотнения, костные дефекты, опухолевый матрикс).

– Межэкспертное согласие оценено по коэффициенту каппы Коэна (κ → 1 для всех признаков).

– В случае расхождений достигнут консенсус в ходе совместного разбора.

Взаимодействие с ChatGPT o3

Версия API: OpenAI ChatGPT o3 на момент тестирования (21 апреля 2025).

Промпт-шаблон: «attached IMG***.jpeg, Задача: Что ты видишь на этой фотографии?»

Параметры вызова: новый сеанс диалога при каждом запросе.

Автоматизация: все запросы отправлялись скриптом на Python с использованием библиотеки requests; между последовательными запросами соблюдалась пауза 3–5 минут для обхода лимитов API.

Политика уточняющих запросов

Для каждого изображения принимался только первый ответ модели. Не задавались последующие уточняющие или корректирующие вопросы, а также не выдавались дополнительные подсказки о содержимом или тематике иллюстрации.

Методы оценки

– Первичная интерпретация: отоскопические изображения оценивал первый эксперт-оториноларинголог (М. В. Комаров), владеющий данными анамнеза пациентов, с фиксацией замечаний и комментариев по каждой иллюстрации.

– Слепая интерпретация: те же изображения независимо оценивал второй эксперт-оториноларинголог (О. И. Гончаров), не знакомый ни с клинической историей пациентов, ни с тем, что анализ выполняет LLM, с фиксацией собственных критических замечаний.

– Консенсусный разбор: совместное обсуждение интерпретаций первого и второго экспертов для выработки окончательного золотого стандарта описания каждого изображения.

– Оценка бинарной точности: консенсусная проверка результатов модели по критерию «норма / патология» с расчетом чувствительности и специфичности.

– Проверка корректности диагноза: консенсусная оценка соответствия диагностической формулировки модели («верность диагноза: да / нет»).

– Согласованность признаков: оценка совпадения перечисленных моделью ключевых морфологических признаков и золотого стандарта экспертов с расчетом коэффициента каппа Коэна (Cohen's κ).

– Субъективная полезность: консенсусная оценка экспертами клинической значимости и удобства использования текстовых описаний модели по 5-балльной шкале Лайкерта (Likert).

Статистическая обработка

Все статистические расчеты выполнялись в среде R и SPSS Statistics 26.0.1.

– Оценка бинарной классификации «норма/патология». Рассчитывались: точность, чувствительность, специфичность, положительная и отрицательная прогностические ценности; 95% доверительные интервалы для пропорций — по точному методу Clopper—Pearson.

– Согласованность признаков и диагнозов. Межэкспертное согласие по ключевым морфологическим признакам определялось с помощью коэффициента каппа Коэна. Согласие экспертной оценки полезности по 5-балльной шкале Лайкерта — с помощью межклассового корреляционного коэффициента.

– Сравнение пропорций. Для проверки статистической значимости различий в доле правильной интерпретации норма/патология и верности диагноза да/нет применялся χ^2-тест (или точный тест Фишера при ожидаемых частотах < 5).

– Непараметрические тесты для шкалы Лайкерта. Для сравнения полезности между более чем двумя группами патологий использовался тест

Краскела—Уоллиса, при парных сравнениях — U-тест Манна—Уитни с поправкой Бонферрони.

– Корреляционный анализ. Для выявления связи между количеством правильно определенных признаков и субъективной полезностью применялся коэффициент Спирмена (Spearman's ρ).

Во всех анализах уровень значимости принимался равным α = 0,05.

Результаты

Общая характеристика интепретаций

На пилотной выборке из 38 отоскопических изображений (3 — норма; 35 — патология) модель продемонстрировала следующие результаты (ALBUM_CLOUD_INTERPRETATION.pdf, рис. 2):

– бинарная точность («норма/патология»): 100% (38/38); 95% CI 90,8–100%;

– чувствительность: 100% (35/35); 95 % CI 90,3–100%;

– специфичность: 100% (3/3); 95% CI 29,2–100%;

– положительная прогностическая ценность (PPV — Positive Predictive Value) и отрицательная прогностическая ценность (NPV — Negative Predictive Value): 100 % (35 / 35 и 3 / 3 соответственно);

– верность постановки диагноза («верность диагноза: да/нет»): 81,6% (31/38); 95% CI 66,6–91,6%;

– соответствие «золотому стандарту» по ключевым морфологическим признакам: 60,5% (23/38); каппа Коэна = 0,82;

– средняя субъективная полезность описаний: M = 4,4 ± 0,6 балла по 5-балльной шкале Лайкерта.

Бинарная точность «норма/патология» и верность постановки диагноза (OTO_comments.pdf, рис. 3)

В табл. 2 приведены результаты бинарной классификации «норма / патология» и доля случаев с корректным диагнозом по золотому стандарту для каждой категории:

– общая точность классификации «норма/патология» : 100% (38/38);

– общая верность постановки диагноза: 81,6% (31/38).

Все случаи были правильно отнесены к «нормальным» или «патологическим». Однако способность модели точно сформулировать клинический диагноз варьировала между категориями: от 0% для опухолей до 100% для нормальных состояний, экссудативного и острого отитов.

Общее резюме применения точного теста Фишера

В нашем исследовании для проверки статистической значимости различий в доле корректных диагностических решений модели ChatGPT 03 между нозологическими группами применялся точный тест Фишера. Этот метод был выбран в силу следующих обстоятельств.

– Во всех категориях число наблюдений не превышало 3–6 случаев. При таких малых объемах χ^2-тест мог давать некорректные *p*-значения из-за низких ожидаемых частот (< 5 в ячейках таблицы).

– Для каждой из 9 групп строилась двумерная таблица «корректный диагноз/некорректный диагноз». Точный тест Фишера позволяет напрямую вычислить вероятность наблюдаемого (или более экстремального) распределения без аппроксимаций.

– Результаты. *p* = 0,012, что значительно ниже выбранного уровня значимости α = 0,05.

– Статистическая значимость указывает на то, что эффективность постановки диагноза моделью существенно различается между группами, главным образом из-за низкой верности формулировки в онкологических случаях.

Рис. 2. QR-код, содержащий ссылку на скачивание файла в формате *.pdf. Название файла ALBUM_CLOUD_INTERPRETATION.pdf. Файл содержит альбом пронумерованных отоскопических иллюстраций, с интерпретацией выполненной LLM. https://disk.yandex.ru/i/CN_NkpDH7wjFhA

Fig. 2. QR code containing a link to download the file in *.pdf format. The file is named ALBUM_CLOUD_INTERPRETATION.pdf. The file contains an album of numbered otoscopic illustrations with interpretations performed by an LLM. https://disk.yandex.ru/i/CN_NkpDH7wjFhA

Рис. 3. QR-код, содержащий ссылку на скачивание файла в формате *.pdf. Название файла OTO_comments.pdf. Файл содержит протокол оценки авторами интерпретации отоскопических иллюстраций, выполненной LLM, с указанием бинарной точности, верности диагноза и критических замечаний М. В. Комарова и О. И. Гончарова https://disk.yandex.ru/i/uLhgiApOwEZuBQ

Fig. 3. QR code containing a link to download the file in *.pdf format. The file is named OTO_comments.pdf. It contains the authors' evaluation protocol of the LLM-generated interpretation of otoscopic illustrations, indicating binary accuracy, diagnostic correctness, and critical remarks by M. V. Komarov and O. I. Goncharov. https://disk.yandex.ru/i/uLhgiApOwEZuBQ

Таблица 2

Бинарная точность «норма/патология» и верность постановки диагноза по нозологическим группам

Table 2

Binary accuracy of "norm/pathology" and the accuracy of diagnosis by nosological groups

Категория	*N*	Бинарная точность «норма/патология» *n* (%)	Верность диагноза *n* (%)
Норма	3	3 (100)	3 (100)
Инородные тела	6	6 (100)	4 (66,7)
Экссудативный отит	6	6 (100)	6 (100)
Посттравматический разрыв перепонки	4	4 (100)	3 (75,0)
Острый средний отит	4	4 (100)	4 (100)
Опухоли среднего уха	3	3 (100)	0 (0)
Ретракционный карман	5	5 (100)	4 (80,0)
Состояние после шунтирования	3	3 (100)	3 (100)
Состояние после радикальной операции	4	4 (100)	4 (100)
Всего	38	38 (100)	31 (81,6)

– Применение точного теста Фишера укрепило достоверность выводов о неоднородности клинической точности ChatGPT o3 и обоснованности необходимости дополнительной «корректировки» prompt-ов и расширения обучающего дата-сета для редких и сложных патологий.

Таким образом, точный тест Фишера стал ключевым статистическим инструментом для объективной оценки различий между малыми группами и обеспечил высокую надежность полученных результатов.

Субъективная оценка клинической полезности

Оценка «человеческой» ценности текстовых описаний модели проводилась двумя экспертами-оториноларингологами по 5-балльной шкале Лайкерта (1 — совершенно бесполезно, 5 — крайне полезно) (табл. 3).

– Средний балл $M = 4{,}4$ свидетельствует о высокой клинической значимости описаний модели.

– Медиана = 5 и межквартильный интервал 4–5 показывают, что более половины оценок равнялись 5 баллам.

– Межэкспертное согласие = 0,82 указывает на «почти полное» согласие экспертов при оценке полезности.

– Отсутствие статистически значимых различий ($p > 0{,}05$) между нозологическими группами говорит об универсальности полезности LLM-описаний для любых категорий отоскопических состояний.

– Распределения субъективных оценок полезности устойчивы: медианы, межквартильные интервалы и плотности оценок не различаются между любыми группами.

– Универсальность восприятия LLM-описаний подтверждает их значение как инструмента для самых разных отоскопических состояний.

– Заключение. Непараметрические тесты показали, что клиническая полезность текстовых описаний ChatGPT o3 не зависит от категории патологии или состояния после вмешательства. Это свидетельствует о стабильном уровне информативности и ценности LLM-интерпретаций в широком спектре диагностических сценариев.

Таблица 3

Субъективная оценка клинической полезности текстовых описаний модели ChatGPT o3

Table 3

Subjective assessment of the clinical usefulness of text descriptions of the ChatGPT o3 model

Параметр	Значение
Средний балл ($M \pm SD$)	4,4 ± 0,6
Медиана (межквартильный интервал)	5 (4–5)
Доля оценок ≥ 4	92%
Межэкспертное согласие	0,82 (95% CI 0,77–0,91)
Тест Краскела—Уоллиса (сравнение полезности между группами)	$\chi^2 = 5{,}43$, $p = 0{,}24$ (различий нет)

Резюме применения ранговой корреляции Спирмена

– Цель анализа состояла в определении существования связи между количеством правильно выявленных моделью ключевых признаков (0–5 на изображение: перфорация, гиперемия, экссудат, тимпаносклероз, холестеатома) и субъективной полезностью ее описаний (1–5 баллов по Лайкерту).

– По каждому из 38 изображений подсчитано число верно обозначенных моделью морфологических признаков. Эксперты выставили оценку полезности текстового описания (1–5 баллов).

– Использован коэффициент ранговой корреляции Спирмена (Spearman's ρ), подходящий для порядковых данных и небольших выборок. Нулевая гипотеза: нет монотонной связи между двумя переменными.

– Результаты: $\rho = 0{,}72$, $p < 0{,}001$. Высокое значение ρ и низкое *p*-значение указывают на статистически значимую сильную положительную корреляцию. С ростом числа правильно определенных моделью признаков эксперты ставили более высокие баллы полезности. Это подтверждает, что точность морфологического анализа напрямую влияет на восприятие клинической ценности LLM-описаний.

– Заключение: наличие значимой положительной связи между точностью распознавания признаков и субъективной оценкой полезности демонстрирует, что повышение качества визуальной интерпретации моделями LLM ведет к росту доверия и практической ценности их выводов.

Время обработки

Время от момента отправки запроса до получения полного текстового ответа модели измерялось вручную автором исследования.

– Среднее время отклика: 30–40 с на одно изображение.

– Диапазон: от 14 до 46 с в зависимости от сложности сцены и нагрузки API.

Такой порядок времени обработки позволяет использовать ChatGPT 03 для первичной скрининг-оценки отоскопических снимков в формате point-of-care, обеспечивая оперативный фидбэк без критических задержек.

Обсуждение

Ключевые результаты

1. Бинарная классификация «норма / патология»:

– модель достигла 100% точности (38/38), без ложноположительных и без ложноотрицательных срабатываний;

– чувствительность и специфичность также составили по 100%, что подтверждает ее способность надежно отделять норму от патологии даже на разнородной выборке из восьми диагностических категорий.

2. Верность постановки клинического диагноза:

– полный клинический диагноз совпал с золотым стандартом в 81,6% случаев (31/38);

– высокие показатели наблюдались для большинства нозологий, однако отмечалось снижение точности при интерпретации иллюстраций опухолей среднего уха (в пределах дифференциального диагноза) и инородных тел наружного слухового прохода.

3. Субъективная полезность текстовых описаний:

– эксперты оценили описания модели в среднем 4,4 ± 0,6 балла по шкале Лайкерта (медиана = 5);

– более 90% оценок составили 4–5 баллов, средний коэффициент согласия каппы Коэна = 0,82 подтверждает «почти полное» межэкспертное согласие.

4. Время обработки:

– среднее время отклика модели на одно изображение составило 30–40 с, что делает ее пригодной для оперативной первичной скрининг-оценки на приеме или в рамках телемедицины.

Вывод: ChatGPT 03 продемонстрировал высокую эффективность и воспроизводимость при комплексном анализе отоскопических изображений, сочетая точную классификацию, качественное описание признаков и приемлемую скорость работы.

Сравнение с существующими подходами

– CNN-модели, хотя и показывают высокую точность в узких задачах (Sn/Sp 92…98%), требуют крупные аннотированные базы и «тонкую» донастройку под каждую патологию.

– LLM-подход предлагает мультинозологическую интерпретацию при неспецифическом запросе «Что ты видишь на этой фотографии?» и естественно-языковую отчетность, что упрощает интеграцию в телемедицинские платформы и образовательные курсы.

Клинические и образовательные импликации

– Телемедицина: оперативный скрининг изображений в первичном звене, предфильтрация «критических» случаев (перфорации, холестеатома).

– Обучение: детальные текстовые описания могут служить интерактивным учебным материалом для студентов и ординаторов, повышая стандартизацию интерпретации отоскопии.

– Электронная медицинская карта: возможность автогенерации диагностических фрагментов отчета, сокращая время документации.

Ограничения исследования

– Объем и однородность выборки: всего 38 кадров в пилотной части анализа. Такая небольшая и клинически однородная выборка ограничивает обобщаемость результатов на другие популяции и условия съемки.

– Несправедливое распределение категорий: значительный перекос в сторону патологических

случаев (35/38) и малая доля нормальных изображений (3/38) могут завышать показатели специфичности и чувствительности модели.

– Ретроспективный дизайн и экспертный консенсус: аннотация золотого стандарта проводилась экспертами по уже имеющимся изображениям, что может создавать эффект «подсматривания» и не отражает слепого независимого тестирования.

– Отсутствие количественных измерений: модель генерирует только текстовые описания признаков без числовых оценок (размер перфорации, степень втяжения, площадь экссудата), что ограничивает применение в научных исследованиях и точных клинических протоколах.

– Субъективность оценки полезности: восприятие клинической ценности текстов оценивалось по шкале Лайкерта двумя экспертами; несмотря на высокое межэкспертное согласие, этот метод остается частично субъективным и требует расширения круга респондентов.

– Зависимость от параметров программного интерфейса и качества изображений: время ответа (30–40 с) и точность модели зависят от версии ChatGPT 03 и качества исходных кадров. Пониженное разрешение, артефакты или нестандартные ракурсы могут снижать воспроизводимость результатов.

Для преодоления этих ограничений необходимы мультицентровая проспективная валидация на значительно расширенной и более сбалансированной выборке, внедрение методов количественного извлечения признаков и расширение числа рецензентов при субъективной оценке.

Перспективные направления

– Мультицентровая проспективная валидация: набор репрезентативной выборки ≥ 5000 отоскопических изображений из различных клиник и устройств; слепое сравнение с независимыми экспертами для проверки переносимости и стабильности показателей.

– Шаблонизация и структурированный вывод: разработка prompt-шаблонов для унифицированного JSON-ответа (текстовый формат обмена данными, основанный на JavaScript), включающего разделенные поля: ключевые признаки, вероятностные оценки, рекомендации по дальнейшим действиям; интеграция в большие электронные базы данных и аналитические конвейеры.

– Мультимодальная интеграция: комбинация текстовой интерпретации LLM с алгоритмами компьютерного зрения для количественного измерения морфометрических параметров (площадь перфорации, объем экссудата, степень втяжения) и данными тимпанометрии/аудиометрии.

– Улучшение онкологической настороженности: построение адаптивных подсказок и дообучение с акцентом на редкие опухолевые и предраковые процессы; создание специализированного справочного дата-сета для онкодиагностики.

– Внедрение в телемедицинские и образовательные платформы: разработка веб/мобильного интерфейса для автоматизированной оценки и обучения с возможностью оперативной консультации в первичном звене и дистанционной подготовки интернов.

– Пользовательский опыт и пояснимость: оценка восприятия описаний клиницистами, исследование влияния объяснений LLM на доверие и скорость принятия решений; внедрение механизма обратной связи.

– Этические и регуляторные аспекты: проработка соответствия требованиям правовых регуляторов, обеспечение безопасности данных пациентов, прозрачность алгоритмических решений перед внедрением в клиническую практику.

Заключение

В ходе пилотного исследования оценены возможности многомодальной LLM ChatGPT 03 для интерпретации отоскопических изображений различных нозологических групп.

1. Модель достигла 100% точности при разделении «норма / патология», без ложных срабатываний, что показывает ее надежность для скрининга.

2. Коэффициент межэкспертного согласия каппа Коэна составил 0,82, что свидетельствует о высокой согласованности с аннотациями экспертов.

3. Модель ChatGPT 03, помимо простого описания структурных изменений, обладает рядом особенностей, делающих ее ценным вспомогательным инструментом для клинициста: LLM безошибочно разделяет «норма/патология» и оперативно маркирует критические случаи для приоритизации, а подробные текстовые описания помогают врачу быстро отличить разные формы отита, перфорации и другие патологии без длительной визуальной оценки. Модель автоматически предлагает клинические рекомендации (туалет под микроскопом, тимпанометрия, микробиологическое исследование, показания к парацентезу или пластике), упрощая планирование лечения и стандартизацию протоколов. Развернутые объяснения можно вставлять в электронную карту, что экономит время на отчетность и служит наглядным пособием для обучения и контроля динамики заболевания. При этом окончательное заключение всегда остается за врачом: LLM лишь облегчает «ориентировочный» скрининг, тогда как учет анамнеза, количественные измерения и данные вспомогательных методов остаются обязательными.

4. Эксперты оценили «человеческие» тексты модели в среднем в 4,4 балла из 5 (межэкспертное

согласие = 0,82), а время отклика 30–40 с позволяет интегрировать LLM в point-of-care и телемедицинские сценарии.

5. Ограниченный объем и однородность выборки, отсутствие количественных метрик и субъективность оценки полезности требуют мультицентровой проспективной валидации на более крупной сбалансированной выборке, разработки структурированного JSON-вывода и инструментов количественного анализа.

Таким образом, результаты демонстрируют высокий потенциал ChatGPT o3 как вспомогательного инструмента для скрининга, обучения и стандартизации интерпретации отоскопии. Для клинического внедрения необходима дальнейшая количественная валидация и шаблонизация выводов.

ЛИТЕРАТУРА/REFERENCES

1. Dubois C, Eigen D, Simon F, Couloigner V, Gormish M, Chalumeau M. Development and validation of a smartphone-based deep-learning-enabled system to detect middle-ear conditions in otoscopic images. *NPJ Digit. Med.* 2024;7:162. https://doi.org/10.1038/s41746-024-01159-9
2. Habib A-R, Kajbafzadeh M, Hasan Z, Wong E, Gunasekera H, Perry C. Artificial intelligence to classify ear disease from otoscopy. *Clin. Otolaryngol.* 2022;47:401-413. https://doi.org/10.1111/coa.13925
3. Lechien JR., Naunheim MR., Maniaci A, Radulesco T, Saibene AM, Chiesa-Estomba CM. et al. Performance and consistency of ChatGPT-4 versus otolaryngologists: a clinical case series. *Otolaryngol. Head Neck Surg.* 2024;0:1-8. https://doi.org/10.1002/ohn.759
4. Livingstone D, Talai AS., Chau J, Forkert ND. Building an Otoscopic screening prototype tool using deep learning. *J. Otolaryngol. Head Neck Surg.* 2019;48:66. https://doi.org/10.1186/s40463-019-0389-9
5. Nam Y., Choi S. J., Shin J., Lee J. Diagnosis of Middle Ear Diseases Based on Convolutional Neural Network. *Comput. Syst. Sci. Eng.* 2023; 46(2): 1519-1532. https://doi.org/10.32604/csse.2023.034192
6. Qu RW., Qureshi U, Petersen G, Lee SC. Diagnostic and management applications of ChatGPT in structured otolaryngology clinical scenarios. *OTO Open.* 2023;7(3):e67. https://doi.org/10.1002/oto2.67
7. Song D, Kim T, Lee Y, Kim J. Image-Based AI Technology for Diagnosing Middle Ear Diseases: A Systematic Review. *J. Clin. Med.* 2023;12:5831. https://doi.org/10.3390/jcm12185831
8. Sundgaard JV., Bray P, Laugesen S, Harte J, Kamide Y, Tanaka C. A Deep Learning Approach for Detecting Otitis Media From Wideband Tympanometry Measurements. *IEEE J. Biomed. Health Inform.* 2022;26(7):2974-2982. https://doi.org/10.1109/JBHI.2022.3159263
9. Tsutsumi K, Goshtasbi K, Risbud A, Khosravi P, Pang J, Lin H. A Web-Based Deep Learning Model. *Otol. Neurotol.* 2021;42(9):e1382-e1388. https://doi.org/10.1097/MAO.0000000000003210
10. Zeng J, Deng W, Yu J, Xiao L, Chen S, Zhang X et al. A deep learning approach to the diagnosis ... using otoscopic images. *Eur. Arch. Otorhinolaryngol.* 2023;280:1621-1627. https://doi.org/10.1007/s00405-022-07632-z
11. Zeng X, Jiang Z, Luo W, Li H, Li H, Li G et al. Efficient and accurate identification of ear diseases using an ensemble deep learning model. *Sci. Rep.* 2021;11:10839. https://doi.org/10.1038/s41598-021-90345-w
12. Шайханова А. К., Поз И. В., Кусембаева Э. А., Асан С. Д., Тлеубаева А. О. Интеграция искусственного интеллекта для обнаружения респираторных заболеваний в программно-аппаратный комплекс «Диагностика на дому». *Вестник КазАТК.* 2024;6(135):272-282.
    Shaikhanova A. K., Poz I. V., Kusembayeva E. A., Asan S. D., Tleubayeva A. O. Integration of artificial intelligence for detection of respiratory diseases in the "Diagnostics at Home" hardware–software complex. *Bulletin of KazATC.* 2024; 6(135):272-282. (In Russ.) https://doi.org/10.52167/1609-1817-2024-135-6-272-282
13. Щепеткин Е. Н., Швалев И. Р., Нохрина Г. Л. Современные подходы к использованию искусственного интеллекта в медицине. *Электронный архив УГЛТУ.* 2024;524-530.
    Shchepetkin E. N., Shvalev I. R., Nokhrina G. L. Modern approaches to the use of artificial intelligence in medicine. *USFEU Electronic Archive.* 2024:524-530. (In Russ.)

Вклад авторов
Все авторы сделали эквивалетный вклад в подготовку публикации.
Contribution of authors
All authors made an equivalent contribution to the preparation of the publication.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.
Conflict of interest: the authors declare no conflict of interest.

Информация об авторах
Комаров Михаил Владимирович — доктор медицинских наук, научный сотрудник отдела патологии наружного, среднего и внутреннего уха, Санкт-Петербургский научно-исследовательский институт уха, горла, носа и речи (190013, Российская Федерация, Санкт-Петербург, Бронницкая ул., д. 9); ассистент кафедры оториноларингологии, Северо-Западный государственный медицинский университет имени И. И. Мечникова; заведующий отделением, врач-оториноларинголог, Городская больница №26 (196240, Российская Федерация, Санкт-Петербург, ул. Костюшко, д. 2); 7_line@mail.ru, https://orcid.org/0000-0003-4471-3603

Гончаров Олег Игоревич — кандидат медицинских наук, ассистент кафедры стоматологии и челюстно-лицевой хирургии, Национальный медицинский исследовательский центр имени В. А. Алмазова (197341, Российская Федерация, Санкт-Петербург, ул. Аккуратова, д. 2); врач-оториноларинголог, Городская больница № 26 (196240, Российская Федерация, Санкт-Петербург, ул. Костюшко, д. 2); https://orcid.org/0000-0003-3738-4944

Федотова Анастасия Александровна — аспирант кафедры оториноларингологии, Северо-Западный государственный медицинский университет имени И. И. Мечникова; врач-оториноларинголог, Городская больница № 26 (196240, Российская Федерация, Санкт-Петербург, ул. Костюшко, д. 2); https://orcid.org/0000-0002-4816-2098

Information about authors

Mikhail V. Komarov — Doctor of Sciences (Med.), Researcher, Department of Pathology of the External, Middle and Inner Ear, Saint Petersburg Research Institute of Ear, Throat, Nose and Speech (9, Bronnitskaya str., Saint Petersburg, Russian Federation, 190013); Assistant, Department of Otolaryngology, Mechnikov North-West State Medical University; Head of Department, Otolaryngologist, City Hospital N 26 (2, Kostyushko str., Saint Petersburg, Russian Federation, 196240); 7_line@mail.ru, https://orcid.org/0000-0003-4471-3603

Oleg I. Goncharov — Candidate of Sciences (Med.), Assistant Professor, Department of Dentistry and Maxillofacial Surgery, Almazov National Medical Research Center (2, Akkuratov str., Saint Petersburg, Russian Federation, 197341); Otolaryngologist, City Hospital N 26 (2, Kostyushko str., Saint Petersburg, Russian Federation, 196240); https://orcid.org/0000-0003-3738-4944

Anastasiya A. Fedotova — Postgraduate Student, Department of Otolaryngology, Mechnikov North-West State Medical University; Otolaryngologist, City Hospital N 26 (2, Kostyushko str., Saint Petersburg, Russian Federation, 196240); https://orcid.org/0000-0002-4816-2098

Поступила / Received 30.03.2025
Поступила после рецензирования / Revised 14.04.2025
Принята в печать / Accepted 06.05.2025