



УДК 61:004.89:616.12-089

DOI 10.17802/2306-1278-2026-15-4-153-163

ПРИМЕНЕНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ИНФОРМАЦИОННОГО ПОИСКА ПРИ ПОДГОТОВКЕ СИСТЕМАТИЧЕСКИХ ОБЗОРОВ И МЕТААНАЛИЗА В СЕРДЕЧНО-СОСУДИСТОЙ ХИРУРГИИ

А.С. Шацкий^{1,2}, Э.М. Деигхеиди¹, С.Е. Масютина^{2,3}, М.Л. Мамалыга¹

¹ Федеральное государственное бюджетное учреждение «Национальный медицинский исследовательский центр сердечно-сосудистой хирургии имени А. Н. Бакулева» Министерства здравоохранения Российской Федерации, пр. Ленинский, 8, к. 7, Москва, Российская Федерация, 117931; ² Малое инновационное предприятие «Лаборатория передовых технологий» при Университете Иннополис, ул. Спортивная, 114, Иннополис, Республика Татарстан, Российская Федерация, 420500; ³ Polytechnic University of Milan, Piazza Leonardo da Vinci, 32, Милан, Италия, 20133

Основные положения

- Большие языковые модели (БЯМ) позволяют автоматизировать первичный скрининг аннотаций при подготовке систематических обзоров, значительно снижая трудоемкость процесса.
- Эффективность работы БЯМ напрямую зависит от качества составления запросов (пром프트-инжиниринг), которые должны четко отражать критерии включения/исключения исследований.
- В статье предложены и проиллюстрированы два подхода к формулировке промптов: «мягкий» (обеспечивает высокую полноту охвата) и «строгий» (снижает число ложноположительных результатов).
- Показана практическая методика загрузки и обработки аннотаций из базы PubMed с использованием БЯМ (на примере DeepSeek, GPT-4o, Perplexity).
- Отмечены ключевые ограничения БЯМ: галлюцинации, чувствительность к формулировкам запросов, смещения в данных обучения, что требует обязательного экспертного контроля результатов.
- Рекомендован двухэтапный подход: сначала мягкий скрининг, затем строгий фильтр, с последующей ручной проверкой релевантных статей.
- Пром프트-инжиниринг рассматривается как практическое искусство, а не универсальный алгоритм, требующее итеративной настройки под конкретные задачи и медицинские области.

Резюме

Большие языковые модели превратились в мощные инструменты автоматизации первичного скрининга аннотаций при проведении систематических обзоров, позволяя существенно снизить ручной труд. В статье представлен всесторонний обзор принципов формирования запросов (пром프트-инжиниринга), демонстрирующий, как традиционные критерии для метаанализа можно преобразовать в четкие инструкции для искусственного интеллекта. На примерах тематических исследований в области хирургии ишемической болезни сердца и врожденных пороков сердца мы иллюстрируем влияние составления промптов на полноту охвата, точность и общую эффективность скрининга литературных источников. Кроме того, рассматриваются типичные ошибки, а также подчеркивается сохраняющаяся необходимость экспертного контроля для минимизации галлюцинаций и смещений, присущих выводам искусственного интеллекта. В итоге систематический пром프트-инжиниринг в сочетании с экспертной оценкой позволяет исследователям в сердечно-сосудистой хирургии оптимизировать поиск источников для метаанализа на базе больших языковых моделей, обеспечивая более быстрый и всесторонний синтез доказательной базы и методов обработки первичного материала.

Ключевые слова

Систематические обзоры • Искусственный интеллект • Сердечно-сосудистая хирургия • Пром프트-инжиниринг • Скрининг источников литературы

Поступила в редакцию: 09.02.2026; поступила после доработки: 23.02.2026; принята к печати: 03.03.2026

Для корреспонденции: Александр Сергеевич Шацкий, alexander.shatsky@gmail.com; адрес: пр. Ленинский, 8, к. 7, Москва, Российская Федерация, 117931

Corresponding author: Alexander S. Shatsky, alexander.shatsky@gmail.com; address: 8, Leninsky Ave., Bld. 7, Moscow, Russian Federation, 117931

USING LARGE LANGUAGE MODELS FOR LITERATURE SEARCH IN CARDIOVASCULAR SURGERY SYSTEMATIC REVIEWS AND META-ANALYSES

A.S. Shatskiy^{1,2}, E.M. Deigheidy¹, S.E. Masyutina^{2,3}, M.L. Mamalyga¹

¹ Federal State Budgetary Institution “National Medical Research Center for Cardiovascular Surgery named after A. N. Bakulev” of the Ministry of Health of the Russian Federation, Leninsky Ave., 8, bld. 7, Moscow, Russian Federation, 117931; ² Small Innovative Enterprise “Laboratory of Advanced Technologies” at Innopolis University, Innopolis city, Sportivnaya St., 114, Republic of Tatarstan, Russian Federation, 420500; ³ Polytechnic University of Milan, Piazza Leonardo da Vinci, 32, Milano (MI), Italy, 20133

Highlights

- Large Language Models (LLMs) enable automation of initial abstract screening in systematic reviews, significantly reducing manual workload.
- The effectiveness of LLMs heavily depends on prompt engineering, which must clearly translate inclusion and exclusion criteria into actionable instructions.
- Two prompt design strategies are presented and illustrated: a “lenient” approach (maximizing recall) and a “strict” approach (reducing false positives).
- A practical workflow for downloading and processing PubMed abstracts using LLMs (e.g., DeepSeek, GPT 4o, Perplexity) is demonstrated.
- Key limitations of LLMs – hallucinations, sensitivity to phrasing, and training data biases — are discussed, emphasizing the need for expert validation.
- A two stage screening strategy is recommended: first a lenient pass, followed by a stricter filter, with subsequent manual verification of relevant articles.
- Prompt engineering is framed as an iterative, domain specific art rather than a one size fits all procedure, requiring continuous refinement for each research question.

Abstract

Large Language Models have evolved into powerful tools for automating the primary screening of articles' abstracts in systematic reviews, enabling a significant reduction in manual labor. The article presents a comprehensive review of prompt engineering principles, demonstrating how traditional meta-analysis criteria can be transformed into clear instructions for artificial intelligence. Using case studies in coronary artery bypass grafting and congenital heart disease surgery, we illustrate the impact of prompt formulation on the comprehensiveness, accuracy, and overall efficiency of literature screening. Furthermore, typical errors are discussed, and the ongoing necessity of expert oversight to minimize hallucinations and biases inherent in artificial intelligence conclusions is emphasized. Ultimately, systematic prompt engineering combined with expert evaluation allows researchers in cardiovascular surgery to optimize the search for meta-analysis sources based on Large Language Models, ensuring a faster and more comprehensive synthesis of evidence and methods for processing primary material.

Keywords

Systematic reviews • Artificial intelligence • Cardiovascular surgery • Prompt engineering • Literature screening

Received: 09.02.2026; **received in revised form:** 23.02.2026; **accepted:** 03.03.2026

Список сокращений

БЯМ – большие языковые модели РКИ – рандомизированное клиническое
ИИ – искусственный интеллект исследование

Введение

Систематические обзоры и метаанализы являются жизненно важными инструментами доказательной медицины, поскольку они синтезируют данные из множества исследований, предоставляя надежные данные об эффективности, безопасности или сравнительной ценности различных медицинских вмешательств [1]. Однако одним из самых трудоемких аспектов проведения система-

тического обзора является скрининг тысяч аннотаций, полученных в результате широкого поиска по базам данных – процесс, требующий много времени и подверженный человеческим ошибкам [2]. Первый этап собора информации может стать существенным препятствием для своевременного и точного синтеза доказательств, особенно по мере увеличения числа публикаций в области сердечно-сосудистой хирургии. Систематические обзоры

обычно начинаются с поиска в таких базах данных, как e-Library.ru, Google Scholar, PubMed, Embase, Scopus, ClinicalTrials.gov, EACTS, ICTRP (ВОЗ) или Cochrane Library, с использованием высокочувствительных запросов [3]. Эти запросы часто очень широки, чтобы не упустить потенциально релевантные исследования. В результате может быть получены тысячи уникальных результатов, и обычно более одного независимого эксперта оценивают названия и аннотации, отбрасывая исследования, которые явно не соответствуют критериям включения [4]. Даже в хорошо скоординированных командах исследователей, этот процесс может быть длительным и подверженным ошибкам из-за непоследовательной интерпретации на каждом из этапов исследования.

Инновации в области искусственного интеллекта (ИИ) привели к разработке нескольких инструментов, помогающих ученым проводить систематические обзоры проще и быстрее [5]. В последнее время появление больших языковых моделей (БЯМ), способных генерировать текст и выполнять сложные задачи текстовой обработки, открыло путь к неожиданным возможностям [6]. Среди них семейство GPT (Generative Pre-trained Transformer) – особенно GPT-3.5 и GPT-4 и семейство DeepSeek привлекли значительное внимание благодаря своей способности интерпретировать и продуцировать язык почти на человеческом уровне. К российским разработкам в этой области относят прежде всего модели Яндекс и ГигаЧат. Системы ИИ могут автоматизировать различные этапы рабочего процесса систематического обзора, включая трудоемкий начальный скрининг названий и аннотаций и исследователи начали использовать БЯМ для выявления релевантных статей для систематических обзоров вместо того, чтобы оставлять эту задачу только командам людей [7].

Однако успешное использование БЯМ для задач скрининга требует большего, чем просто предоставление этим алгоритмам примерного круга интересующих вопросов. Модели ИИ нуждаются в точном наборе инструкций для эффективной работы. Таким образом, правильный промпт-инжиниринг, то есть создание четких, содержательных инструкций, является ключом к получению желаемого результата, и недавние публикации подчеркнули актуальность этих навыков для любой медицинской специальности, будь то кардиология, анестезиология или кардиохирургия [8].

В этой статье кратко рассматривается, что такое БЯМ, различия между подходами, и как эти концепции применяются к скринингу биомедицинской литературы. После установления этих основ мы предоставим небольшое руководство по написанию запросов – от более мягких до более строгих стилей, – которые направляют ИИ к включению

или отклонению аннотаций статей на основе заданных критериев.

Мы надеемся, что исследователи в области сердечно-сосудистой хирургии и не только, смогут использовать эти методики для улучшения своих систематических обзоров, поддержания высоких стандартов научных исследований и, в конечном итоге, практике принятия медицинских решений, основанных на доказательствах, в здравоохранении в целом.

Что такое большие языковые модели

Эволюция БЯМ, началась со статистических языковых моделей в 1990-х гг., которые полагались на вероятностные методы для предсказания последовательностей слов на основе цепей Маркова. Им на смену в начале 2010-х гг. пришли нейросетевые языковые модели, использующие глубокое обучение для лучшего улавливания взаимосвязей. Современная эра БЯМ, вероятно, началась с серии GPT от OpenAI, ознаменовав переход к крупномасштабным моделям на основе архитектуры Transformer, способным генерировать текст, похожий на человеческий [9]. Архитектура Transformer, представленная в 2017 г., сыграла решающую роль в реализации этих достижений, эффективно справляясь с длинными зависимостями и превосходя подходы на основе рекуррентных сетей. «Трансформеры» используют механизм под названием self-attention для оценки важности разных слов в контексте, что позволяет модели обрабатывать длинные зависимости гораздо эффективнее, чем более ранние модели [10]. К 2022 г. OpenAI усовершенствовала свою модель до GPT-3.5, улучшив связность и снизив склонность к генерации не по теме или ошибочных утверждений. Эта версия стала широко доступной через ChatGPT (версия 3.5), значительно повысив беглость ведения диалога и следование сложным инструкциям. В 2023 г. OpenAI выпустила GPT-4, также доступную в оптимизированной форме (GPT-4o), что еще больше расширило границы управления контекстом и логической последовательности, хотя ее точная архитектура и количество параметров раскрываются. GPT-4 показала особые перспективы в специализированных областях, таких как биомедицинские исследования, где такие задачи, как обобщение научных статей или скрининг аннотаций для систематических обзоров, выигрывают от ее продвинутых возможностей обработки естественного языка [11].

Примерно в то же время другие организации вступили в гонку БЯМ, предложив альтернативы моделям OpenAI. Google представила Gemini, которая, как сообщается, включает мультимодальные рассуждения и напрямую конкурирует с GPT-4. Anthropic разработала Claude, сосредото-

чившись на безопасной и прозрачной генерации языка, в то время как Alibaba представила Qwen, подчеркивая возможности многоязычной поддержки и специализации в предметных областях. Дополнительные модели, такие как DeepSeek, Ян-декс и ГигаЧат отражают тенденцию расширения архитектур БЯМ, причем каждая новая система заявляет об уникальных преимуществах – будь то безопасность, масштаб, поддержка нескольких языков или снижение галлюцинаций. Хотя многие из этих новых моделей еще не прошли широкого тестирования для биомедицинских приложений, они предоставляют альтернативные пути, особенно для задач автоматизированного скрининга [12]. Обучаясь на триллионах токенов (слов или субъединиц слов), БЯМ развивают удивительно широкое понимание языка, стиля, фактических знаний и даже некоторых способностей к рассуждению. БЯМ могут выполнять такие задачи, как обобщение текста, перевод языков, ответы на вопросы и классификация контента – и все это без явного программирования или тонкой настройки для каждой конкретной задачи, а в результате их крупномасштабного обучения.

Ограничения использования больших языковых моделей в медицине

Несмотря на эти сильные стороны, системы ИИ также демонстрируют ключевые ограничения, которые могут повлиять на их полезность для специализированных задач. Одна заметная проблема связана с галлюцинациями, когда модель уверенно предоставляет фактически неверную информацию [13]. Например, она может сослаться на рандомизированное контролируемое исследование в журнале, которого на самом деле не существует, с вымышленными авторами и выводами. В контексте систематических обзоров такие галлюцинации могут быть особенно вредны, если исследователи ошибочно доверяют этим вымышленным деталям, тем самым ставя под угрозу надежность отбора исследований или извлечения данных.

Следует отметить, что на практике сами аннотации статей могут упускать важные детали (например, точную продолжительность наблюдения или методы рандомизации). В таких случаях перевод критериев включения/исключения в конкретные запросы, может столкнуться с трудностями, требующими запасных стратегий (например, пометки машиной неоднозначных аннотаций статей для проверки человеком).

Другая проблема – это чувствительность к форме запроса, когда незначительное изменение формулировки инструкций для различных моделей ИИ, может привести к заметно разным результатам [14]. В качестве иллюстрации, изменение инструкции с «Принять, если участники – X» на

«Отклонить, если участники – не X» может привести к непредвиденным последствиям для данной конкретной модели, даже несмотря на схожую внутреннюю логику.

Критерии выбора источников поиска релевантной информации

Определение того, какую часть статьи использовать для скрининга, является критическим решением в систематических обзорах и имеет существенные последствия как для точности, так и для эффективности.

Традиционно команды исследователей начинают с рассмотрения только названий, которые можно быстро оценить, чтобы отбросить явно нерелевантные источники – например, те, которые сосредоточены на другой популяции пациентов или не связаны с исследуемым вмешательством [5]. Мы и другие авторы успешно использовали названия во многих работах, потому что названия эффективно обобщают тему статьи. Однако, названия часто содержат мало деталей о методологии и могут привести к большому количеству ложных срабатываний при скрининге, переводя многие нерелевантные элементы на следующий этап.

Еще один, более информативный вариант – это аннотации, часто называемые в обиходе российских научных работников «абстрактами» статей, которые широко доступны через основные бесплатные базы данных (например, e-Library/РИНЦ, Medline/PubMed и Google Scholar) без каких-либо ограничений по оплате, в отличие от полных текстов, которые могут требовать подписки [15]. Аннотации обычно обобщают популяцию исследования, вмешательства, первичные конечные точки и, во многих случаях, методологические детали, такие как рандомизация или продолжительность наблюдения. Используя этот краткое, но всеобъемлющее резюме исследования, читатели и рецензенты (а теперь и ИИ) могут принять обоснованное решение о соответствии статьи критериям. Скрининг аннотаций представляет собой золотую середину, которая балансирует глубину информации с практичностью: текст достаточно краток для быстрого анализа алгоритмами ИИ, но обычно достаточно подробен, чтобы уловить ключевые критерии включения или исключения. Кроме того, аннотации стандартизированы во многих биомедицинских отчетах, что делает их относительно единообразными.

Потенциальным расширением скрининга является рассмотрение полных текстов статей, которые могут устранить неоднозначность, позволяя рецензенту или БЯМ получить доступ ко всем методологическим деталям и деталям исходов. Этот подход, однако, создает значительные логистические и вычислительные проблемы. Во-первых, по-

лучение полных текстов – особенно для большого числа результатов поиска – может потребовать дополнительных подписок или межбиблиотечного абонемента (например, системы Российской государственной библиотеки), а многие базы данных не размещают полные статьи. Во-вторых, в случае с БЯМ загрузка длинных документов в системы ИИ может быть вычислительно затратной, потенциально превышая лимиты токенов (запросов, слов), которые в продвинутых версиях часто предоставляются на платной основе. Обработка полного текста также увеличивает время вывода на статью, что для тысяч источников может стать довольно дорогим [16].

Для авторов систематических обзоров, стремящихся внедрить БЯМ, принятие этапа скрининга на основе аннотаций, представляется эффективной отправной точкой, хорошо согласующейся с существующими рабочими процессами, бюджетными ограничениями и легко извлекаемым характером резюме статей в основных библиографических базах данных, а в дальнейшем позволяет искать пути доступа к полному тексту, даже если она содержится в платных источниках, таких как платформы ScienceDirect, Ovid и Wiley. Отдельно следует упомянуть систему Oxford Academic, которая предоставляет выход к большому массиву академических журналов, включая ключевые в кардиохирургии ЕС – European Heart Journal, European Journal of Cardio-Thoracic Surgery (EJCTS), Interactive CardioVascular and Thoracic Surgery (ICVTS), Journal of Surgical Case Reports и другие, на платной и бесплатной основе. Доступ к части данных журналов можно получить также и через сайт Европейской ассоциации сердечно-сосудистых хирургов (www.EACTS.org).

Наиболее доступным является поиск и загрузка файлов предварительно отобранных аннотаций статей или только их названий через систему PubMed. Для этого необходимо зайти на страницу <https://pubmed.ncbi.nlm.nih.gov/advanced/>. Затем, набрать запрос в поисковую строку, либо ввести комбинацию слов в поле Query box в виде команды

с логическими операторами OR (или), AND (и), в зависимости от вашего опыта использования данной системы (рис. 1).

После поиска, на экране появится множество названий найденных источников. Для создания файла с массивом найденных «абстрактов», нужно нажать кнопку Save, в поле Selection выбрать All results, а в поле Format – Abstract (text), как изображено на рис. 2, и создать файл (Create file).

Файл сохранится в виде текста (с расширением .txt), который будет необходимо загрузить в чат-бот ИИ перед созданием соответствующих запросов, описанных далее (рис. 3).

Создание запросов в системах больших языковых моделей

Запрос (или промпт, от англ. prompt – запрос) – это, по сути, набор инструкций или контекстной информации, которую пользователь предоставляет БЯМ, чтобы направлять ее вывод. В простейшем случае промпт может быть просто быстрым вопросом, введенным в строку поиска, – но он также может быть очень подробным, определяя стиль, тон, структуру и даже точный формат желаемого ответа [17]. Вообще говоря, хорошо структурированные запросы обычно содержат четыре ключевых элемента:

- 1. **Инструкция:** определяет задачу, которую должна выполнить модель, или поведение, ко-

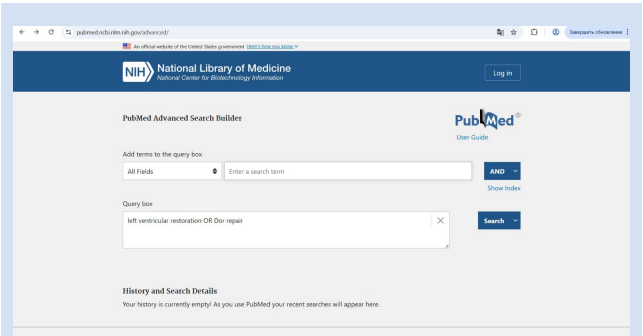


Рисунок 1. Диалоговое окно Национальной медицинской библиотеки США с введенным запросом по теме исследования
Figure 1. Search interface of the U.S. National Library of Medicine showing a query entered for the research topic

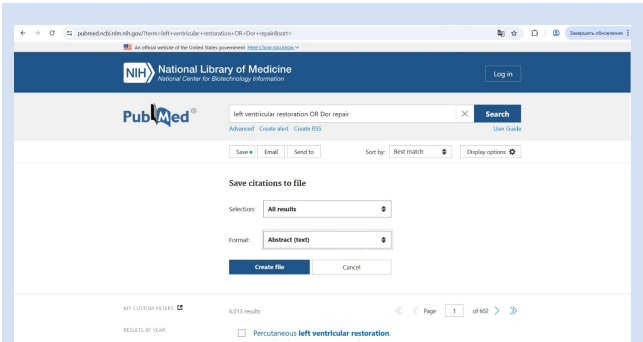


Рисунок 2. Сохранение файла данных с текстовым массивом аннотаций статей для последующего анализа
Figure 2. Saving a data file with a text array of article abstracts for subsequent analysis

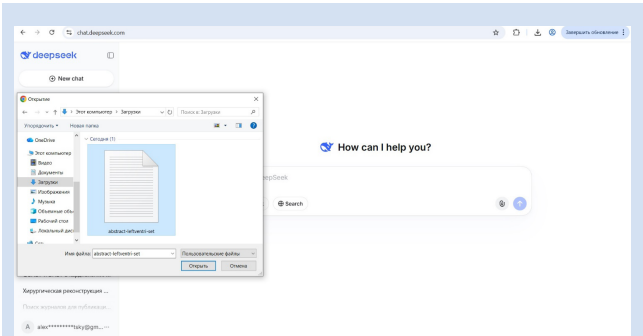


Рисунок 3. Диалоговое окно большой языковой модели Deep Seek в момент загрузки файла для анализа
Figure 3. Uploading a file to the DeepSeek large language model for subsequent analysis

торое она должна принять (например: «Обобщите следующие аннотации по теме»)

2. **Контекст:** предлагает фоновые детали, специфичные для предметной области, которые помогают модели давать более точные и релевантные ответы (например, разъяснение, что в данном случае ИИ – это помощник, помогающий в систематическом обзоре по хирургии ишемической болезни сердца).

3. **Входные данные:** конкретный текст или вопрос, который модель должна обработать, чтобы она точно знала, какой контент анализировать (например, «только аннотации, описывающие РКИ или когортные исследования»).

4. **Индикатор вывода:** указывает желаемый формат или тип ответа – будь то короткий суммирующий абзац, список источников или более подробное объяснение в виде обобщения полученной информации.

Пользователи могут комбинировать эти элементы, чтобы согласовать ответы БЯМ со своими конкретными целями. В конкретном случае скрининга литературы, запрос будет включать контекст, критерии соответствия и инструкции по принятию решений. Структурированной основой, часто используемой в клинических исследованиях для определения поисковых вопросов и критериев включения/исключения, является аббревиатура PICO – Популяция (Population), Вмешательство (Intervention), Сравнение (Comparison), Исход (Outcome) [18].

Например, можно начать с контекста:

– «Необходимо помочь в проведении систематического обзора по оценке результатов операций хирургического лечения постинфарктных аневризм сердца путем сравнения отдаленных результатов двух типов коррекций по методике Дор и Кули», тем самым определяя конкретную предметную область.

Затем пользователь описывает ключевые требования, ставя задачу и определяя входные данные (критерии включения/исключения – согласно PICO):

– «Взрослая популяция (≥ 18 лет), наличие постинфарктной аневризмы, виды вмешательств операции Дор и Кули, дизайн РКИ или когортное, минимальное наблюдение шесть месяцев, оценка выживаемости, язык исследования английский»

и, наконец, точно инструктируем модель, как отвечать, например:

«Если аннотация удовлетворяет этим критериям включения, ответ «ПРИНЯТЬ»; в противном случае ответ «ОТКЛОНИТЬ». Вывести список принятых аннотаций статей в виде таблицы. Без дополнительного текста».

Такая структура обеспечит ситуацию, когда ИИ точно обучен, какие признаки искать, где их найти и как форматировать классификацию (результат).

Структура промпта имеет решающее значение, потому что БЯМ изначально не «знают», как поступить с каждым запросом. Вместо этого они полагаются на внешние директивы для структурирования и интерпретации данных. Хорошо разработанный запрос способствует ясным и последовательным ответам, тогда как неясные или неполные промпты могут привести к путанице или несвязанному выводу. В конечном счете, промпты функционируют как важный мост между обширными предварительно усвоенными знаниями модели и целями пользователя, превращая универсальную систему ИИ в целенаправленный и более надежный инструмент для таких задач, как обобщение результатов исследований или выявление релевантных статей в систематическом обзоре.

Установка критериев отбора

Важным аспектом промпт-инжиниринга является определение того, насколько строгой должна быть модель ИИ при оценке аннотаций.

• **Разрешительный (или «мягкий») промпт** инструктирует БЯМ принимать статью, если она явно не нарушает предопределенное требование. Например, в промпте может быть указано, что единственным основанием для отклонения являются явные противоречия, такие как исключительно детская популяция, если критерии включения включают возраст > 18 лет. Этот подход обычно дает высокую полноту охвата, гарантируя, что потенциально релевантные исследования не будут случайно исключены; однако он также может привести к увеличению количества ложноположительных случаев, тем самым требуя больше проверок человеком.

• **Строгий промпт** предписывает БЯМ отклонять статью, если она явно не соответствует каждому критерию включения. Эта стратегия может снизить количество ложноположительных случаев, требуя четких критериев, таких как явное упоминание пациентов-взрослых, дизайна РКИ и указанной продолжительности наблюдения, но может непреднамеренно исключить релевантные исследования, аннотации которых не детализируют каждый требуемый элемент.

Учитывая, что многие систематические обзоры отдают приоритет чувствительности (т.е. минимизации ложноотрицательных случаев), но также стремятся не допустить большого количества статей для проверки на последующем этапе скрининга полного текста, используют двухэтапный подход, при котором в первой итерации использует мягкий промпт, а для второго «прохода» применяют более строгий фильтр, что может обеспечить тщательный

скрининг без чрезмерного ручного последующего перебора.

Чтобы показать, как эти стратегии могут выглядеть на практике, можно начать с мягкого промпта, такого как следующий:

«Необходимо помочь в проведении систематического обзора по вопросу аритмических осложнений при ишемической кардиомиопатии, сравнивая группы с проведенной реваскуляризацией миокарда плюс радиочастотная абляция против группы только с реваскуляризацией миокарда».

Включаем: ретроспективные и проспективные исследования с наблюдением не менее шести месяцев у взрослых (≥ 18 лет) и не старше 80 лет с ишемической кардиомиопатией без рандомизации.

Если аннотация статьи предполагает или не противоречит этим критериям, ответ 'ПРИНЯТЬ'. Ответ 'ОТКЛОНИТЬ' только если она явно нарушает любое требование: возраст менее 18 и более 80 лет, наблюдение менее шести месяцев, упоминания рандомизации. Выведи слово 'ПРИНЯТЬ' или 'ОТКЛОНИТЬ'. Результат – в виде таблицы».

Этот дизайн обеспечивает высокую полноту охвата, потому что модель будет принимать исследования, если они окончательно не «проваливают» правило включения.

В отличие от этого, строгий промпт может выглядеть так:

«Необходимо помочь в проведении систематического обзора по вопросу аритмических осложнений при ишемической кардиомиопатии, сравнивая группы с проведенной реваскуляризацией миокарда плюс радиочастотная абляция против группы только с реваскуляризацией миокарда».

Включаем ретроспективные и проспективные исследования с наблюдением не менее шести месяцев у взрослых (≥ 18 лет) и не старше 80 лет с ишемической кардиомиопатией без рандомизации.

Если какой-либо критерий отсутствует или неясен, ответ 'ОТКЛОНИТЬ'.

В противном случае ответ 'ПРИНЯТЬ'. Результат – в виде таблицы.

Без дополнительного текста».

Здесь отсутствие явного упоминания требуемого элемента ведет к исключению аннотации, тем самым снижая ложноположительные случаи, но увеличивая риск упустить те из них, которые все еще могут быть релевантны, если недостающая деталь подтвердится позже уже в полном тексте.

Хотя продвинутые БЯМ, такие как GPT-4 или GPT-4o, предлагают собственные пути для полуавтоматизации скрининга литературы, несколько

проблем могут снизить их производительность, если их не устранить специальным образом.

Типичные примеры ошибочных запросов и некорректных результатов

- **Чрезмерно длинные запросы:** одна частая проблема связана с использованием излишне длинных промптов, в которых чрезмерная фоновая информация размывает фокус модели на основных правилах классификации [19]. Современные модели, безусловно, могут обрабатывать большие объемы текста, но они могут отклониться от своей основной задачи, если инструкции «похоронены» в длинных или второстепенных материалах.

- **Противоречивые или неоднозначные критерии:** другое препятствие возникает, когда промпты содержат противоречивые или неоднозначные критерии, например, инструктируя модель отклонять любую аннотацию без деталей исследуемой группы, но одновременно смягчая запрос и не устраняя неопределенность какие собственно критерии нужно использовать [20]. Такие конфликтующие инструкции часто приводят к непоследовательным или непредсказуемым ответам.

- **Галлюцинации и неявные выводы:** БЯМ иногда могут галлюцинировать, уверенно выводя характеристики, которые фактически не указаны, например, помечая исследование как «рандомизированное» просто потому, что оно описано как «проспективное». Включение в запрос включения только аннотаций, в которых явно упоминают только «РКИ», «рандомизированные исследования» или другие однозначные термины, помогает уменьшить долю таких ошибочных ответов.

- **Смещение в выводах** – еще одна потенциальная область беспокойства. БЯМ обычно обучаются на больших данных, собранных из интернета, которые могут непропорционально представлять исследования из англоязычных или высокодоходных регионов, имеющих финансовую возможность публиковаться в рейтинговых журналах. В результате модель может систематически «научиться» отдавать предпочтение определенным типам публикаций или непреднамеренно недооценивать исследования из недостаточно представленных языков или точек зрения. Например, аннотации, написанные не на английском языке, или те, в которых сообщается о редких заболеваниях или группах, могут получать менее точные классификации. Этот перекокс может внести тонкие, но существенные искажения в процесс скрининга, потенциально упуская ценные доказательства или создавая неполную картину доступных исследований [21].

Борьба с этими рисками включает несколько стратегий:

- **Необходимость итеративного процесса:** промпт-инжиниринг редко совершенствуется с первого раза, поэтому итеративный процесс тестирования, доработки и повторной проверки промптов на небольшом валидационном наборе часто очень важен для выявления запросов, где возникают ложноположительные и ложноотрицательные случаи, и соответствующей калибровки инструкций [20]. При этом бывает важно учитывать, что при отборе аннотаций на английском языке, бывает необходимо сопоставить терминологию в русскоязычном научном сообществе с терминами на языке оригинала и опробовать оба варианта (например, вместо АКШ – аортокоронарное шунтирование, написать в запросе CABG – coronary artery bypass grafting и т. д.).

- **Незаменимость человеческого надзора:** несмотря на значительные достижения в возможностях ИИ, человеческий надзор остается незаменимым, особенно для пограничных случаев, связанных с неясными аннотациями или необычными дизайнами исследований. Хотя модель может отфильтровать большие массивы нерелевантных статей, экспертная оценка важна для проверки точности окончательных решений и обеспечения того, чтобы никакие критические доказательства не были ошибочно исключены. Хотя представленный здесь пример сосредоточен на хирургии ишемической болезни сердца, аналогичные стратегии применимы к другим областям, таким как кардиология или другие биомедицинские исследования. Адаптация запросов к различным медицинским специальностям, очевидно, может представлять специальные проблемы. Аннотации по трансплантации сердца могут требовать точных определений биомаркеров (например, типов HLA-антигенов), в то время как исследования по функциональной диагностике могут зависеть от различия терминов визуализации (например, режимов МРТ и МСКТ), и их придется решать с помощью экспертов в предметной области и уточнения запросов, так как неверный ввод критериев приведет к неверному результату.

Исследователи могут сами отслеживать систематические ошибки. Периодические «точечные проверки» решений модели в сочетании с опытом использования нескольких систем ИИ с поиском как российских, так и англоязычных источников, позволяют выявлять и находить те БЯМ, которые наиболее хорошо подходят под собственные задачи. Наконец, сам промпт-инжиниринг можно использовать, чтобы побудить модель ИИ активно проверять, достаточно ли представлены соответ-

ствующие темы, тем самым снижая вероятность ошибочной классификации, вызванной смещением, если вы регулярно используете один и тот же чат-бот [22].

Заключение

Продвинутые БЯМ обладают большим потенциалом для снижения рабочей нагрузки по скринингу аннотаций в систематических обзорах. Однако достижение точных, последовательных и прозрачных результатов зависит от написания эффективных запросов, которые переводят ваши критерии в недвусмысленные, проверяемые инструкции. В этой статье было показано, как стратегии «мягких» и «строгих» промптов различаются по полноте охвата и точности, обсуждались подходы к классификации и были предоставлены конкретные примеры того, как эти методологии могут быть применены в тематическом исследовании в области хирургии ишемической болезни сердца.

Хотя скрининг с поддержкой ИИ не заменит полностью человеческое суждение, он позволяет командам врачей-исследователей сосредоточить свое время на задачах более творческих, таких как критическая оценка полных текстов статей, извлечение данных и их синтез.

Создание промптов – это скорее практическое искусство, а не универсальная процедура и не язык программирования. Общие принципы: явное указание критериев принятия/отклонения, короткие и последовательные форматы вывода и тщательное тестирование на валидационном наборе – широко применимы в медицинских областях. Некоторые исследователи предпочтут высоко «мягкий промпт», чтобы обеспечить полноту охвата, близкую к 100%, в то время как другие могут полагаться на строгий подход, чтобы ограничить ложноположительные случаи.

Поскольку модели БЯМ продолжают развиваться, а специализированные виды ИИ для конкретных медицинских областей, например, Foreseemed и Pathway.md, набирают популярность и даже интегрируются с системами предоставления медицинской информации, такими как UpToDate и, вероятно, когда-нибудь, подобные инструменты станут стандартом. Тем не менее, бдительность в отношении смещений, галлюцинаций или противоречивых выводов остается важной, и окончательный надзор со стороны обученных исследователей необходим для сохранения целостности доказательной медицины. Это, в свою очередь, может ускорить научный прогресс и способствовать более своевременному, всестороннему пониманию постоянно расширяющейся сферы специализированной медицинской литературы.

Конфликт интересов

А.С. Шацкий заявляет об отсутствии конфликта интересов. Э.М. Деигхеиди заявляет об отсутствии конфликта интересов. С.Е. Масютина заявляет об отсутствии конфликта интересов. М.Л. Мамалыга заявляет об отсутствии конфликта интересов.

Финансирование

Авторы заявляют об отсутствии финансирования исследования.

Информация об авторах

Шацкий Александр Сергеевич, Master of Science (University of Bologna), кандидат медицинских наук докторант-соискатель института коронарной и сосудистой хирургии федерального государственного бюджетного учреждения «Национальный медицинский исследовательский центр сердечно-сосудистой хирургии имени А.Н. Бакулева» Министерства здравоохранения Российской Федерации, Москва, Российская Федерация; заместитель директора по научной работе Малого инновационного предприятия «Лаборатория передовых технологий» при Университете Иннополис, Иннополис, Республика Татарстан, Российская Федерация; **ORCID** 0009-0003-3654-1618

Деигхеиди Эхаб Махруз, кандидат медицинских наук докторант-соискатель института кардиохирургии им. В.И. Бураковского федерального государственного бюджетного учреждения «Национальный медицинский исследовательский центр сердечно-сосудистой хирургии имени А.Н. Бакулева» Министерства здравоохранения Российской Федерации, Москва, Российская Федерация; **ORCID** 0007-2929-4021

Масютина Стефания Евгеньевна, Master of Science (Politecnico di Milano), ведущий научный сотрудник Малого инновационного предприятия «Лаборатория передовых технологий» при Университете Иннополис, Иннополис, Республика Татарстан, Российская Федерация; научный сотрудник Polytechnic University of Milan, Милан, Италия; **ORCID** 0009-0005-9668-1211

Мамалыга Максим Леонидович, доктор медицинских наук ведущий научный сотрудник отдела хирургического лечения ишемической болезни сердца института коронарной и сосудистой хирургии федерального государственного бюджетного учреждения «Национальный медицинский исследовательский центр сердечно-сосудистой хирургии имени А. Н. Бакулева» Министерства здравоохранения Российской Федерации, Москва, Российская Федерация; **ORCID** 0000-0002-7444-9930

Author Information Form

Shatskiy Alexander S., Master of Science (University of Bologna), PhD, Doctoral Candidate at the Institute of Coronary and Vascular Surgery of the Federal State Budgetary Institution “National Medical Research Center for Cardiovascular Surgery named after A.N. Bakulev” of the Ministry of Health of the Russian Federation, Moscow, Russian Federation; Deputy Director for Research at the Small Innovative Enterprise “Laboratory of Advanced Technologies” at Innopolis University, Innopolis city, Republic of Tatarstan, Russian Federation; **ORCID** 0009-0003-3654-1618

Deigheidy Ehab M., PhD, Doctoral Candidate at the V.I. Burakovsky Institute of Cardiac Surgery, Federal State Budgetary Institution “National Medical Research Center for Cardiovascular Surgery named after A. N. Bakulev” of the Ministry of Health of the Russian Federation, Moscow, Russian Federation; **ORCID** 0009-0007-2929-4021

Masyutina Stefaniya E., Master of Science (Politecnico di Milano), Leading Researcher at the Small Innovative Enterprise “Laboratory of Advanced Technologies” at Innopolis University, Innopolis city, Republic of Tatarstan, Russian Federation; Researcher at Polytechnic University of Milan, Milano, Italy; **ORCID** 0009-0005-9668-1211

Mamalyga Maxim L., PhD, MD, Leading Researcher at the Department of Surgical Treatment of Coronary Heart Disease, Institute of Coronary and Vascular Surgery, Federal State Budgetary Institution “National Medical Research Center for Cardiovascular Surgery named after A. N. Bakulev” of the Ministry of Health of the Russian Federation, Moscow, Russian Federation; **ORCID** 0000-0002-7444-9930

Вклад авторов в статью

ШАС – вклад в концепцию и дизайн исследования, анализ данных исследования, написание и корректировка статьи, утверждение окончательной версии для публикации, полная ответственность за содержание

ДЭМ – вклад в концепцию исследования, анализ данных исследования, корректировка статьи, утверждение окончательной версии для публикации, полная ответственность за содержание.

МСЕ – вклад в концепцию и дизайн исследования, получение и интерпретация данных исследования, корректировка статьи, утверждение окончательной версии для публикации, полная ответственность за содержание

ММЛ – вклад в концепцию и дизайн исследования, корректировка статьи, утверждение окончательной версии для публикации, полная ответственность за содержание

Author Contribution Statement

ShAS – contribution to the concept and design of the study, data analysis, manuscript writing, editing, approval of the final version, fully responsible for the content

DEM – contribution to the concept of the study, data analysis, editing, approval of the final version, fully responsible for the content

MSE – contribution to the concept and design of the study, data collection and interpretation, editing, approval of the final version, fully responsible for the content

MML – contribution to the concept and design of the study, editing, approval of the final version, fully responsible for the content

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Parums, D. V. Review Articles, Systematic Reviews, Meta-Analysis, and the Updated Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 Guidelines // *Medical Science Monitor*. 2021. Vol. 27. P. e934475.
2. Waffenschmidt S., Knelangen M., Sieben W., Bühn S., Pieper D. Single Screening versus Conventional Double Screening for Study Selection in Systematic Reviews: A Methodological Systematic Review // *BMC Medical Research Methodology*. 2019. Vol. 19. Art. No. 132.
3. Bramer W. M., Rethlefsen M. L., Kleijnen J., Franco O. H. Optimal Database Combinations for Literature Searches in Systematic Reviews: A Prospective Exploratory Study // *Systematic Reviews*. 2017. Vol. 6. Art. No. 245.
4. Cooper C., Booth A., Varley-Campbell J., Britten N., Garside R. Defining the Process to Literature Searching in Systematic Reviews: A Literature Review of Guidance and Supporting Studies // *BMC Medical Research Methodology*. 2018. Vol. 18. Art. No. 85.
5. Chai K. E. K., Lines R. L. J., Gucciardi D. F., Ng L. Research Screener: A Machine Learning Tool to Semi-Automate Abstract Screening for Systematic Reviews // *Systematic Reviews*. 2021. Vol. 10. Art. No. 93.
6. Gates A., Johnson C., Hartling L. Technology-Assisted Title and Abstract Screening for Systematic Reviews: A Retrospective Evaluation of the Abstrackr Machine Learning Tool // *Systematic Reviews*. 2018. Vol. 7. Art. No. 45.
7. Khraisha Q., Put S., Kappenberg J., Warritch A., Hadfield K. Can Large Language Models Replace Humans in the Systematic Review Process? Evaluating GPT-4's Efficacy in Screening and Extracting Data from Peer-Reviewed and Grey Literature in Multiple Languages. arXiv:2310.17526 [Preprint]. 2023.
8. Zaghir J., Naguib M., Bjelogrić M., Névél A., Tannier X., Lovis C. Prompt Engineering Paradigms for Medical Applications: Scoping Review // *Journal of Medical Internet Research*. 2024. Vol. 26. P. e60501.
9. Wang Z., Chu Z., Doan T. V., Ni S., Yang M., Zhang W. History, Development, and Principles of Large Language Models: An Introductory Survey // *AI Ethics*. 2024. P. 1–17.
10. Shiri F. M., Perumal T., Mustapha N., Mohamed R. A Comprehensive Overview and Comparative Analysis on Deep Learning Models. arXiv:2305.17473 [Preprint]. 2023.
11. Gue C. C. Y., Rahim N. D. A., Rojas-Carabali W., Agrawal R., Palvannan R. K., Abisheganaden J., Yip W. F. Evaluating the OpenAI's GPT-3.5 Turbo's Performance in Extracting Information from Scientific Articles on Diabetic Retinopathy // *Systematic Reviews*. 2024. Vol. 13. Art. No. 135.
12. Wang J., Yang Z., Yao Z., Yu H. JMLR: Joint Medical LLM and Retrieval Training for Enhancing Reasoning and Professional Question Answering Capability. arXiv:2402.17887 [Preprint]. 2024.
13. Huang L., Yu W., Ma W., Zhong W., Feng Z., Wang H., Chen Q., Peng W., Feng X., Qin B., et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions // *ACM Transactions on Information Systems*. 2025. Vol. 43. P. 1–55.
14. Sclar M., Choi Y., Tsvetkov Y., Suhr A. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting. arXiv:2310.11324 [Preprint]. 2023.
15. Saloojee H., Pettifor J. M. Maximizing Access and Minimizing Barriers to Research in Low- and Middle-Income Countries: Open Access and Health Equity // *Global Health, Science and Practice*. 2022. Vol. 10, No. 1. P. e2100403.
16. Ye A., Maiti A., Schmidt M., Pedersen S. J. A Hybrid Semi-Automated Workflow for Systematic and Literature Review Processes with Large Language Model Analysis // *Future Internet*. 2024. Vol. 16, No. 5. P. 167.
17. Liu P., Yuan W., Fu J., Jiang Z., Hayashi H., Neubig G. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing // *ACM Computing Surveys*. 2023. Vol. 55, No. 9. P. 1–35.
18. Eriksen M. B., Frandsen T. F. The Impact of Patient, Intervention, Comparison, Outcome (PICO) as a Search Strategy Tool on Literature Search Quality: A Systematic Review // *Journal of the Medical Library Association*. 2018. Vol. 106, No. 4. P. 420–431.
19. Kusano G., Akimoto K., Takeoka K. Are Longer Prompts Always Better? Prompt Selection in Large Language Models for Recommendation Systems. arXiv:2412.14454 [Preprint]. 2024.
20. Heston T. F., Khun C. Prompt Engineering in Medical Education // *International Medical Education*. 2023. Vol. 2, No. 4. P. 198–205.
21. Lee J., Hicke Y., Yu R., Brooks C., Kizilcec R. F. The Life Cycle of Large Language Models in Education: A Framework for Understanding Sources of Bias // *British Journal of Educational Technology*. 2024. Vol. 55, No. 4. P. 1982–2002.
22. Huang D., Bu Q., Zhang J., Xie X., Chen J., Cui H. Bias Testing and Mitigation in LLM-Based Code Generation. arXiv:2309.14345 [Preprint]. 2023.

Для цитирования: Шацкий А.С., Деигхеиди Э.М., Масюткина С.Е., Мамалыга М.Л. Применение больших языковых моделей для информационного поиска при подготовке систематических обзоров и метаанализа в сердечно-сосудистой хирургии. *Комплексные проблемы сердечно-сосудистых заболеваний*. 2026;15(4): 153-163. DOI: 10.17802/2306-1278-2026-15-4-153-163

To cite: Shatskiy A.S., Deigheidy E.M., Masyutina S.E., Mamalyga M.L. Using large language models for literature search in cardiovascular surgery systematic reviews and meta-analyses. *Complex Issues of Cardiovascular Diseases*. 2026;15(4): 153-163. DOI: 10.17802/2306-1278-2026-15-4-153-163

Приложение 1

Дополнительные примеры запросов для практического использования

Ниже приведены дополнительные примеры, которые исследователи могут адаптировать к своим собственным критериям. Эти запросы мягкие, обеспечивают высокую полноту охвата. Они также улавливают суть критериев включения, в связи с тем, что аннотации могут быть неполными. Исследователи могут сначала просмотреть все статьи, которые БЯМ классифицирует как подходящие, а затем отсеять их вручную или применить второй, более строгий скрининг, при желании. Данные примеры опробованы для ИИ моделей DeepSeek, Perplexity и GPT-4o. По нашему субъективному опыту, наиболее релевантные ответы получены при использовании GPT-4o, но он не доступен напрямую в РФ, поэтому, второе место в этом ряду мы бы присудили китайской модели DeepSeek. Единственным плюсом использования Perplexity является то, что он частично может производить поиск путем прямого доступа к базам Elsevier или PubMed, хотя количество выдаваемых им источников ограничено 10 (в бесплатных версиях), в остальном его ответы кажутся нам менее структурированными. Для обработки больших баз аннотаций, рекомендовано использовать несколько моделей ИИ, в связи с незначительными отличиями в логике и ответах систем, а затем следует просматривать все релевантные статьи для их оценки медицинской командой.

Мягкий промпт без обучающего примера (использование ксеноперикардального кондуита в хирургическом лечении врожденных пороков сердца)

Предварительно, произведен поиск в PubMed, по ключевым словам: Ross procedure AND congenital heart diseases. Получено примерно 27 тыс. аннотаций. Файл с аннотациями загружен в чат-бот. Структура запроса представлена ниже.

«Вы – помощник, помогающий в систематическом обзоре по использованию ксеноперикардального кондуита в хирургическом лечении врожденных пороков сердца. Мы будем решать, следует ли каждую аннотацию ПРИНЯТЬ или ОТКЛОНИТЬ на основе следующих критериев:

Популяция: Пациенты – дети (возраст менее 18 лет) с любым врожденными пороками сердца. Если аннотация не упоминает возраст или явно не описывает детские популяции, не исключай.

Вмешательство: Должно включать хирургическое лечение с использованием ксеноперикардального кондуита – xenopericardial conduit. Если ксеноперикардальный конduit упоминается частично, не исключай.

Сравнение: Группа, которая включает любое вмешательство при пороках конотрункуса. Если не указан метод хирургического лечения, не исключай.

Исходы: Отдаленная выживаемость. Если аннотация не указывает исходы, но упоминает ксеноперикардальный конduit, не исключай.

Дизайн исследования: Должно быть проспективное или подразумевать случайное распределение. Если неясно, не исключай.

Наблюдение: Более 6 месяцев. Если не указано или неясно, не исключай.

Если ни один критерий не нарушен явно, «ПРИНЯТЬ». Если какой-либо критерий явно нарушен (взрослая популяция, наблюдение < 6 месяцев), «ОТКЛОНИТЬ». Без дополнительного текста. Результат в виде списка».

Мягкий промпт с одним обучающим примером (лечение диабетической стопы при сахарном диабете 2 типа)

Предварительно произведен поиск в PubMed, по ключевым словам: diabetes mellitus AND SGLT2. Получено примерно 7 тыс. результатов. Файл .txt загружен в чат-бот.

«Вы – ИИ-помощник, помогающий в систематическом обзоре по лечению сахарного диабета 2 типа (СД2) с использованием ингибиторов SGLT2. Вы будете решать, следует ли каждую новую аннотацию ПРИНЯТЬ или ОТКЛОНИТЬ на основе следующих критериев:

Популяция: Пациенты-взрослые (≥ 18 лет) с сахарным диабетом 2 типа. Если аннотация не указывает возраст, не исключай.

Вмешательство: Должно включать хотя бы один ингибитор SGLT2 (канаглифлозин, эмпаглифлозин, дапаглифлозин). Если аннотация содержит другой ингибитор SGLT2, не исключай.

Сравнение: включает плацебо или любой другой противодиабетический лекарственный препарат. Если не указано, но не противоречит, не исключай.

Исходы: Должны сообщать об изменениях в состоянии периферического кровообращения при диабетической стопе. Если аннотация подразумевает улучшение гликемических показателей без явного упоминания показателей периферического кровотока, не исключай.

Дизайн исследования: Должно быть РКИ или подразумевать случайное распределение. Если неясно, не исключай.

Наблюдение: Минимум 3 месяца. Если не указано или неясно, не исключай.

Ниже приведен пример аннотации, соответствующей всем критериям (помечена как ПРИНЯТЬ):

Аннотация: [Пользователю необходимо вставить пример аннотации путем прикрепления файла в чат-бот].

Используя приведенный выше пример в качестве ориентира, классифицируй следующие аннотации на «ПРИНЯТЬ» или «ОТКЛОНИТЬ». Если ни один критерий не нарушен явно, ответ «ПРИНЯТЬ». Если какой-либо критерий явно нарушен (исследование только на детях, нерандомизированный дизайн, длительность наблюдения < 3 месяцев), ответ «ОТКЛОНИТЬ». Не давай дополнительных объяснений. Результат в виде списка».

Данные выкладки не являются образцом языка программирования и не требуют обязательного строгого порядка использования, они могут быть применены частями в зависимости от потребности в выявлении конкретных данных или по этапам в процесс диалога с машиной, каждый раз уточняя предыдущий запрос.