ChatGPT

Комплексное исследование: архитектуры, протоколы и технологии для AI-терапевта

Категория 1: Технические архитектуры гибридных систем

- **Сценарный диалоговый менеджмент (Script-Based Policy):** Недавняя работа предложила интегрировать LLM в детерминированный сценарий с конечными состояниями, где «скрипт» определяет терапевтический ход беседы, а LLM генерирует реплики в рамках этих правил [1] [2]. Такой подход позволяет боту строго соблюдать заданную структуру терапии и делать ход беседы прозрачным и отслеживаемым.
- **Script-Strategy Aligned Generation (SSAG):** В развитии идеи сценариев, метод **SSAG** сочетает гибкость LLM с опорой на ключевые элементы терапевтического сценария. Исследование 2024 года показало, что чатбот, **выравненный со сценариями экспертов**, превосходит как чисто LLM-модель, так и чисто rule-based бота — достигается баланс между гибкостью диалога и следованием принципам терапии [3] [4]. SSAG снизил трудоемкость скриптов: LLM доучивается лишь на основных стратегиях, что упрощает разработку.
- **Конечные автоматы и явное управление состояниями:** Практические реализации уже появились. Например, в блоге LangChain описан AI-терапевт **Sonia**, моделирующий сессию КПТ как **конечный автомат из 8 стадий** по протоколу Бека [5]. Для каждой стадии строится своя цепочка агентов и промптов, а переходы между стадиями управляются комбинацией правил (напр. ограничение по числу сообщений) и дополнительных LLM-запросов на проверку достижения целей этапа [6] [7]. Такой гибрид обеспечивает структуру при сохранении персонализации.
- **DSL и оркестрация функций:** Индустриальные фреймворки предлагают **Domain-Specific Language** для сценариев. Например, **OpenAI Function Calling** позволяет вызывать функции (действия) из LLM, фактически превращая намерения пользователя в вызов определенного сценария. Это, по сути, встроенный механизм управляемого диалога. Похожим путем пошла Rasa в системе **CALM (Conversational AI with Language Models)**: LLM используется для преобразования естественного языка пользователя в **команды на DSL**, по которым детерминированный движок выполняет логику диалога [8] [9]. Такой подход объединяет мощь понимания LLM с надежностью rule-based исполнения.

Категория 2: Терапевтические протоколы и их реализация в AI

- **Когнитивно-поведенческая терапия (КПТ):** КПТ славится структурированностью – типичная сессия проходит через ~8 этапов (моделирование ситуации, определение мыслей, эмоциональная оценка и т.д.) [5]. Это делает КПТ удобной для алгоритмизации. Чатбот **Woebot** – ранний пример, реализующий техники КПТ. В РКИ 2-недельного использования Woebot у молодых людей были достигнуты значимые снижения симптомов депрессии по сравнению с контрольной группой [10]. В другом исследовании Woebot показал **не худшие** результаты по снижению тревожности, чем общение с живым терапевтом, что свидетельствует о потенциале таких систем.

- **Мотивационное интервьюирование (МИ):** Это метод, помогающий повысить мотивацию к изменениям (например, отказ от вредных привычек). Недавно представлен полностью генеративный чатбот-консультант по МИ для курильщиков. Он был создан совместно с клиническими специалистами и продемонстрировал обнадеживающие результаты: после беседы с ботом уверенность участников в способности бросить курить выросла в среднем на **1,7 пункта из 10**, а автоматический анализ показал, что **98% реплик бота соответствовали техникам МИ** (что даже выше, чем у человеческих консультантов) [11] [12] . Пользователи также отметили высокий (хоть и чуть ниже человеческого) уровень эмпатии от бота. Это говорит о том, что современные LLM способны соблюдать принципы МИ и позитивно влиять на установки клиента.
- **"Internal Family Systems" (IFS, терапия внутренних суб-личностей):** Хотя научных публикаций об автоматизации IFS пока мало, энтузиасты уже пытаются реализовать чатботов, помогающих клиенту исследовать свои «части личности». Существуют прототипы наподобие *IFS Buddy* и других приложений, которые ведут пользователя через процесс самотерапии, задавая вопросы от лица воображаемого терапевта. Эти системы опираются на сценарии IFS (внутренние диалоги с «частями») и служат скорее вспомогательными инструментами для самостоятельной практики. Открытых исследований эффективности пока нет, но подобные инициативы показывают, что даже такие сложные протоколы, как IFS, могут быть частично формализованы для AI.
- **Другие методы:** Помимо КПТ и МИ, активно изучается реализация и других подходов. Например, протоколы **диалектической поведенческой терапии (DBT)** для управления эмоциями или **терапии принятия и ответственности (ACT)** – их элементы (упражнения майндфулнес, когнитивные переоценки) можно прописать в сценариях, а LLM – использовать для эмпатичного ведения. Пока эти направления менее разработаны, но прогресс LLM-систем предполагает, что широкий спектр доказательных методик (экспозиция при фобиях, техники позитивной психологии, коучинг навыков и пр.) могут быть адаптированы в конверсационный формат. Ключевое – тесное сотрудничество с психотерапевтами при создании таких ботов, чтобы сохранить клиническую достоверность.

Категория 3: Безопасность, риск-менеджмент и Guardrails

- **Фильтрация вредного контента:** AI-терапевт должен строго избегать высказываний, усугубляющих состояние пользователя. Поэтому встраиваются многоуровневые **guardrails** – от словарей запретов до моделей-модераторов. Например, **NVIDIA NeMo Guardrails** предлагает язык Colang для задания правил: разработчик прописывает нежелательные темы или фразы, и движок блокирует или переформулирует ответы LLM, выходящие за границы [13] [14] . Открытые библиотеки вроде **Guardrails AI** (от Shreya Rajpal) также позволяют задавать структурные ограничения на ответы (формат, тон) и проверять их перед выводом. Для терапевтического бота особенно важно отсекать советы нанести вред себе/другим, триггерные утверждения, сексуально некорректные фразы и т.п.
- **Обработка кризисных ситуаций (суицидальная идеация):** Бот должен **распознавать сигналы о кризисе** и действовать по протоколу – предоставить контакты кризисной службы, настоять на привлечении людей. Однако исследования показывают, что многие существующие приложения справляются с этим неудовлетворительно. В недавнем испытании 29 различных ментальных чатботов с суицидальными сценариями **ни один бот не выполнил все критерии адекватной реакции**; лишь ~52% дали хотя бы минимально приемлемый ответ, а остальные 48% ответили явно недостаточно [15] [16] . Частые проблемы – бот не сообщает информацию о службе помощи и не улавливает контекст. Этот случай подчёркивает необходимость улучшения: современные рекомендации предполагают встроить в бота детектор по ключевым словам/

эмоциональному тону, который при срабатывании *немедленно* переводит диалог на скрипт кризисной поддержки (либо подключает живого специалиста).

- **Защита от манипуляций и сбоев модели:** Враждебные или нетипичные запросы пользователя (например, попытки обмануть бота, **prompt injection** атаки) могут вывести LLM из роли. Поэтому архитектура должна ограничивать «творчество» модели рамками сценария. Один из подходов – **дисциплинировать выход**: например, требовать от модели ответа строго в формате JSON или короткой фразы командного DSL, который затем парсится детерминированно. Такой метод снижает шанс, что модель начнёт "фантазировать". Также применяются вторичные модели-мониторы: т.н. *referee agent* просматривает черновой ответ LLM и блокирует/правит его, если видит отклонение от терапевтической этики или заданной роли. OpenAI реализует похожую идею через **GPT-решильдинг** (модель-критик, дорабатывающая ответ другой модели).
- **Отслеживание эмоционального состояния и риска:** AI может пассивно мониторить тональность пользователя – например, повышение негативной лексики, признаки паники или агрессии. При обнаружении таких сигналов система может автоматически **усилить фильтры** (более мягкие ответы, более простые фразы) либо переключить сценарий на успокаивающие техники. Если же происходит эскалация (упоминания о нежелании жить, о насилии и т.п.), бот обязан выдать **тревожное оповещение**: показать пользователю ресурсы для неотложной помощи, попросить обратиться к врачу, а в некоторых случаях – если известны экстренные контакты – сообщить их. В клинических условиях подобные автоматические алерты могут перенаправлять сессию к дежурному психотерапевту.

Категория 4: Парсинг пользователя и анализ вводимых данных

- **Профилирование и слоты:** Чтобы вести терапию, бот должен **"понять" пользователя** – его проблему, эмоции, установки. Для этого из сообщений извлекаются ключевые **слоты**: например, симптоматика (депрессия, тревога), упоминания важных персон, когнитивные искажения, цели терапии. Современные LLM уже умеют распознавать подобные детали по контексту диалога. Например, если пользователь говорит *"Мне нечего ждать от будущего"*, модель может пометить это как *"безнадежность"* и *"когнитивное искажение: катастрофизация"*. Исследования подтверждают, что NLP-модели способны автоматически классифицировать когнитивные искажения в тексте не хуже экспертов-психологов [17] [18]. Это открывает путь к чатботам, которые будут в реальном времени подмечать мышление клиента (напр., "чёрно-белое мышление", "чтение мыслей") и мягко направлять его к более здоровым интерпретациям.
- **Отслеживание состояния (Dialog State Tracking):** Помимо долгосрочного профиля, есть задача поддерживать *диалоговое состояние* – что уже обсуждено, к каким выводам пришли, какую технику применяют сейчас. Терапевтический бот может вести **"память о сессии"**: напр., хранить, какие темы пользователь уже поднимал (семья, работа, страхи), какие "домашние задания" давались, какие улучшения отметились. Специализированные NLP-пайплайны могут извлекать из сообщений пользователя индикаторы прогресса: улучшилось настроение или нет, какие препятствия всплыли. Также практикуется авто-анализ тональности (sentiment analysis) каждого сообщения – чтобы бот реагировал на изменение эмоционального фона собеседника.
- **Карты убеждений и целей:** Интересный подход – построение своего рода *когнитивной карты* клиента. На основе его рассказов модель может составить список основных негативных убеждений (например: "Я ничего не добьюсь в жизни") и проверять их в ходе диалога. При работе по КПТ бот бы возвращался к этим убеждениям, когда уместно, чтобы оспорить или переосмыслить их совместно с пользователем. Технически это можно реализовать через извлечение **ключевых убеждений** (комбинация NER и кластеризации

тем) и хранение их как отдельных сущностей. Затем, когда бот генерирует ответ, ему можно подсказать в промпте: *“учти, что у пользователя есть убеждение X, постарайся адресовать его”*. Подобные механизмы пока экспериментальны, но они могут приблизить процесс к настоящей терапии, где терапевт держит в уме “карту” проблем клиента.

- **Оценка прогресса:** AI может автоматически оценивать успехи по ходу общения. Например, если в начале разговора пользователь самоуничижительно отзывался о себе, а через 4 недели общения начал формулировать мысли более позитивно, модель заметит снижение негативных слов и вывод “установки смягчаются”. Также боты способны администрировать психологические опросники прямо в чате – задавая по одному вопросу (PHQ-9, GAD-7 и т.д.) и **парся ответы в баллы**. Это делается либо шаблонным разбором (по ключевым словам ответа) либо прямым вопросом к модели: *“Определи балл ответа по шкале 0-3”*. Таким образом, система оценивает симптомы в динамике, что помогает адаптировать план терапии.

Категория 5: Специфика родительского отчуждения и высококонфликтных семей

- **Родительское отчуждение (PA) – контекст:** PA – это ситуация, когда один из родителей систематически настраивает ребёнка против другого после развода [19]. Это сложный, **многофакторный феномен**, часто требующий вмешательства суда и специалистов (психологов, социальных работников) [20]. Последствия PA для психики ребенка тяжелые: исследования отмечают, что пережившие отчуждение дети во взрослом возрасте чаще страдают от депрессии, низкой самооценки, трудностей в отношениях [21] [22]. Проблема усугубляется тем, что сами семьи, вовлеченные в PA, редко признают наличие проблемы – каждый родитель склонен винить другого, а ребенок находится под влиянием, искажая свои воспоминания о “таргетном” родителе.
- **Терапевтические подходы при PA:** Обычная семейная терапия не всегда эффективна, как отмечают эксперты [23] [24]. Разработаны специальные вмешательства: **реунификационная терапия** (reunification therapy) – разновидность семейной терапии, где основной фокус на восстановлении отношений с отчужденным родителем. В тяжелых случаях суд может временно передать опеку другому родителю и назначить интенсивную программу восстановления связи (известны программы типа *Family Bridges* по Ворошаку). В целом, терапия включает когнитивную работу с ребенком (развенчание ложных убеждений о “плохом” родителе), коучинг для родителей (прекращение отчуждающего поведения) и постепенное возобновление контактов.
- **Роль технологий:** Хотя прямых исследований применения AI именно к PA мало, **цифровые инструменты** уже помогают в смежных областях. Например, существуют онлайн-платформы для ко-перинга (совместного воспитания после развода), снижающие конфликтность за счет структурированного общения. NLP можно применить для **выявления эмоционально окрашенных, враждебных высказываний** в переписке между родителями – по сути, детектировать язык вражды или попытки очернить второго родителя. Отдельные энтузиасты разрабатывают модели для анализа судебных документов и сообщений на признаки отчуждения [25] [26], хотя это пока на уровне прототипов. Возможен и чатбот-помощник для “таргетных” родителей: он мог бы давать эмоциональную поддержку и обучать стратегиям взаимодействия с ребенком, основанным на принципах КПТ (например, помогать отрабатывать терпеливое и последовательное поведение, несмотря на враждебность ребенка). Такой бот действовал бы скорее как **коуч для родителя**, помогая справиться со стрессом и не усугублять ситуацию.
- **Этические моменты:** Вмешательство AI в такие деликатные семейные конфликты требует особой осторожности. Неправильный совет может навредить ребенку или усилить

конфронтацию. Поэтому любой AI-инструмент в области PA должен разрабатываться при участии семейных психологов и юристов. Возможно, наиболее реалистичное применение – *психообразование* через чатбот: объяснение родителям сущности отчуждения, совет по поведению (в духе руководств от экспертов, напр. Бернета или Ворошака), проверка понимания. Это может повысить доступность знаний о PA, но терапевтические функции AI здесь, вероятно, останутся ограниченными и обязательно дополняться человеческим контролем.

Категория 6: Фреймворки и инструменты для разработки

- **LangChain и LangGraph:** Для Python-экосистемы одним из стандартных инструментов стал **LangChain**, предлагающий модули памяти, менеджеры диалога и даже визуальное конструирование цепочек. В 2024 году представлена надстройка **LangGraph**, которая облегчает создание диалогов как **графа состояний** (узлы – функции или подсказки, ветви – условия перехода) [27] . Разработчик может явно задать этапы беседы и логику переходов, а узлы графа могут использовать LLM для генерации текста. LangChain также обеспечивает интеграцию с памятью (разделяемый контекст между узлами), что упрощает построение сложных терапевтических сценариев с длительным контекстом.
- **Rasa Conversational AI with LLM (CALM):** Open-source фреймворк Rasa традиционно использовался для intent-ориентированных ботов. Новая архитектура **Rasa CALM** совмещает подход интентов с LLM: модель берет на себя интерпретацию пользовательского ввода и генерацию действий. В работе Rasa описано, что LLM **переводит пользовательское сообщение в доменно-специфичный язык (DSL)**, а далее встроенный диалоговый менеджер исполняет этот сценарий детерминированно [8] [9] . Таким образом, Rasa CALM позволяет использовать мощь GPT-4 (или другой LLM) внутри проверенного pipeline'а: разработчик задаёт логику на понятном DSL, а LLM гибко обрабатывает естественные фразы пользователя. На практике это даёт значительное упрощение разработки сложных ботов по сравнению с чисто intent-моделью, и при этом сохраняет надежность исполнения бизнес-логики.
- **NeMo Guardrails:** От NVIDIA доступна библиотека **NeMo Guardrails**, которая особенно полезна для безопасных приложений. Она содержит язык **Colang** – по сути, **DSL для правил и диалоговых схем**. Разработчик прописывает на Colang паттерны (например: *если пользователь спросил X, а бот ответил Y, то сделать Z*; *если фраза содержит оскорбление – извиниться и сменить тему*). Guardrails acts как прослойка: перехватывает запросы/ответы LLM и модифицирует их по заданным правилам [13] [28] . Для AI-терапевта NeMo Guardrails может использоваться, например, чтобы внедрить "этики": гарантировать, что бот не даёт советов по медикаментам, не обсуждает запрещенные темы или всегда выполняет проверки настроения каждые N сообщений.
- **Xatkit – чатбот как конечный автомат:** Проект **Xatkit** предоставляет Java-фреймворк с декларативным языком для описания ботов. Разработчик определяет **намерения пользователя, события и действия**, связывая их в **диалоговое дерево/автомат** с помощью fluent-интерфейса на Java [29] [30] . Xatkit поддерживает интеграцию собственных NLP-моделей для классификации интентов, а с выходом LLM можно использовать генерацию для ответов. Этот инструмент фокусируется на четком определении логики разговоров, что созвучно задачам AI-терапии – где нужен контроль. Кроме того, Xatkit позволяет подключать разные платформы (Telegram, веб и т.д.) и потому удобен для прототипирования.
- **Другие:** На GitHub появляется множество репозиториев, полезных для создания терапевтических ботов. Например, **Hugging Face Transformers** предоставляет готовые модели эмоционального анализа и даже диалоговые модели, которые можно адаптировать под терапию. Библиотеки вроде **OpenAI Moderator** или **Perspective API**

могут быть встроены для автоматического цензурирования токсичных ответов. Для долгосрочной памяти популярны vector store решения – **FAISS** или **Chroma DB**, через которые бот может делать *retrieval* прошлых сессий. Наконец, существуют специализированные SDK: скажем, **Google Dialogflow CX** поддерживает гибридные агенты (правила + LLM) и визуальное моделирование состояний – это может быть альтернативой для enterprise-разработки AI-терапевтов.

Категория 7: Оценка эффективности и валидация

- **Клинические испытания (RCT):** Золотой стандарт – проверка AI-терапевтов в рандомизированных контролируемых исследованиях. В **2025 году опубликовано первое в мире RCT** полностью генеративного терапевтического чатбота (*Therabot*) [31]. В испытании (106 участников с депрессией, тревожным или пищевым расстройством) чатбот-терапия длилась 4 недели. Результаты впечатляют: у пациентов с депрессией снижение симптомов составило ~51% от исходного уровня, при тревоге – ~31%, что сопоставимо с результатами очной терапии [32]. Участники с риском расстройств пищевого поведения тоже показали улучшение (среднее снижение озабоченности весом на 19%, что значительно лучше контроля) [33]. Кроме того, люди сообщили о **высоком уровне доверия** к боту – они могли открыто делиться переживаниями, доверяя Therabot почти так же, как живому психологу [34]. Это говорит о формировании терапевтического альянса, ключевого для успеха терапии.
- **Метрики терапевтического альянса:** Эффективность AI-терапии измеряется не только снижением симптомов, но и качеством взаимодействия. Используются опросники наподобие **WAI (Working Alliance Inventory)** адаптированные для чатбота. В вышеупомянутом RCT участники оценили эмпатичность и понимание со стороны бота довольно высоко, хотя и чуть ниже, чем у человеческих терапевтов [12]. Некоторые исследования фиксируют, что пользователи ценят анонимность бота и отсутствие осуждения, что **парадоксально усиливает откровенность** – люди чаще раскрываются о болезненных вещах, зная, что перед ними программа. Это позитивно сказывается на альянсе. Тем не менее, недостаток живого человеческого контакта тоже ощущается: часть участников отмечает, что бот “не может по-настоящему понять, потому что сам не человек”. Такие субъективные показатели нужно учитывать при доработке системы (например, усиливать эмоциональную отзывчивость бота).
- **Долгосрочные исходы и отказоустойчивость:** Пока большинство испытаний охватывают краткосрочный период (несколько недель использования). Важный вопрос – **стойкость эффектов**: сохраняется ли улучшение после прекращения общения с ботом? Будут ли пользователи продолжать применять навыки, которым научил AI? Для ответа требуются длительные исследования на месяцы и более. Ещё один аспект – процент **отсева пользователей**: насколько люди привыкают или, наоборот, быстро забрасывают бота? Если, скажем, из 100 начавших только 20 продолжают диалог через месяц, это сигнал о проблемах с мотивацией или UX системы. Метрики engagement’а (DAU/MAU, средняя длительность сессии, частота пользовательских сообщений) становятся прокси для “привлекательности” терапевтического бота. Некоторые опубликованные пилоты сообщают о высокой вовлеченности: например, **80% пользователей добровольно вернулись за повторной сессией** в течение недели после первого чата – что обнадеживает. Однако без стандартизированных сравнений пока сложно судить, насколько AI-терапевт удерживает клиента относительно живого.
- **Качественная обратная связь:** Помимо чисел, исследователи собирают отзывы пользователей. Из них видно, что многие ценят **доступность 24/7** – можно написать боту в любой момент кризиса. Также отмечают структурированность: “бот помог разложить мои мысли по полочкам”. Негативные отзывы касаются недостаточной гибкости (“иногда

отвечал шаблонно") и отсутствия реакции на невербальные сигналы (что естественно, ведь бот видит только текст). Эти качественные данные помогают улучшать дизайн диалогов: добавлять больше вариативности ответов, предусматривать новые сценарии, где раньше бот пасовал. Например, если пользователи жалуются, что бот избегает отвечать на прямые вопросы ("А что мне делать со своей проблемой?"), разработчики могут добавить соответствующий скрипт с более прямыми рекомендациями или разъяснениями, почему бот не может дать готового совета.

Категория 8: Этика и регулирование

- **Руководства по ответственному развитию:** Понимая риски, эксперты публикуют рекомендации. В 2024 году группа исследователей сформулировала **принципы ответственной разработки клинических LLM** [35] [36]. Ключевые пункты: (1) **участие профильных специалистов** – психотерапевты должны активно участвовать в создании бота и задавать его функциональность; (2) **привязка к доказательным методам** – AI должен применять только проверенные терапевтические практики и действовать в рамках установленного протокола, без отсебятины; (3) **инспектируемость** – поведение бота должно быть прозрачно для аудита, то есть каждое решение желательно объяснимо через переходы между известными состояниями; (4) **безопасность превыше всего** – предусмотреть механизмы предотвращения вреда, реагирования на кризис, защиты данных. Эти принципы во многом и легли в основу архитектур, о которых речь шла выше (сценарный контроль, логирование всех шагов и т.д.).
- **Конфиденциальность и данные:** Медицинская информация крайне чувствительна. Чатбот-терапевт неизбежно будет работать с **персональными данными о здоровье**, поэтому обязан соблюдать стандарты конфиденциальности (в США – HIPAA, в Европе – GDPR и национальные законы). На практике это означает шифрование переписки, хранение данных на защищенных серверах, получение информированного согласия у пользователя на обработку данных. Многие компании изначально позиционируют свои приложения как *не медицинские*, чтобы избежать строгого регулирования, но тренд идет к тому, что успешные AI-терапевты будут сертифицироваться как **медицинские изделия**. Например, **Woebot Health** уже запустил клинические испытания для регистрации своего чатбота как **Software as a Medical Device (SaMD)** с одобрения регулирующих органов [37]. Это потребует доказать эффективность и безопасность подобно лекарственному средству.
- **Регуляторные органы и стандарты:** Всемирная организация здравоохранения (ВОЗ) выпустила в 2024 г. руководство по этике и управлению ИИ в здравоохранении, включая большие мультимодальные модели [38] [39]. В нём подчёркивается необходимость **прозрачности, оценки рисков, участия сообщества и недопущения дискриминации**. ВОЗ рекомендует правительствам развивать специальные регуляторные механизмы для AI в медицине, вплоть до лицензирования и постмаркетингового надзора [40] [41]. В Англии действует **Evidence Standards Framework (NICE)**, устанавливающий уровни доказательности для цифровых терапий – AI-терапевт должен соответствовать определённым критериям клинической эффективности, экономической ценности и т.д. В США FDA выпустило руководство по программам поддержки клинических решений (CDS), где оговорено, в каких случаях ПО считается медицинским и требует одобрения. AI-терапевт, дающий конкретные рекомендации пациенту, скорее всего подпадает под эту категорию.
- **Этические дилеммы:** Есть и **философские вопросы**: допустимо ли, чтобы машина заменила живого терапевта? Не пострадает ли сострадание и аутентичность помощи? Как избежать зависимости пользователей от чатбота? Эксперты по этике указывают, что AI должен рассматриваться *как дополнение, а не замена* человеку [42]. Правильная метафора – "терапевт в кармане", который поддерживает между сессиями с психологом, либо

помогает тем, у кого нет доступа к терапии. Также требуется честность с пользователем: он должен знать, что говорит с AI, а не человеком. Прозрачность алгоритмов – еще один принцип: желательно объяснять, на основе чего бот дает тот или иной совет (например: “Я заметил, что вы используете обобщения, это когнитивное искажение, поэтому предлагаю взглянуть на факты…”). И наконец, вопрос ответственности: если AI даст неверный совет и человек пострадает, кто виноват? Пока закон однозначно не отвечает, поэтому разработчики стремятся **ограничить сферу применения** ботов (например, не допускать самодиагностики серьезных состояний, всегда рекомендовать обратиться к врачу при тяжёлых симптомах и т.п.). В ближайшие годы ожидается формирование отраслевых этических кодексов и, вероятно, сертификация специалистов по AI-терапии, которые будут отвечать за корректность таких систем.

Категория 9: Память и контекст в диалоге

- **Краткосрочная память (текущий сеанс):** LLM уже умеют удерживать довольно большой контекст (несколько тысяч токенов), но терапевтическая сессия может быть длительной. Обычно применяют **резюме-контекст**: периодически длинный чат сворачивают в краткое резюме ключевых точек, которое остается в prompt. Например, после 10 сообщений бот формирует: “мы обсудили X, Y; клиент чувствует Z; договорились попробовать W”. Это сжимает контекст и позволяет модели не забыть важных деталей. Такой подход реализован в ряде фреймворков памяти LangChain (ConversationBufferWindowMemory, Summarizer). Он же помогает, если пользователь “прыгает” между темами – резюме фиксирует каждую нить разговора.
- **Долгосрочная память:** В терапии важно учитывать историю взаимоотношений с клиентом, прогресс сессий. Поэтому внедряется **персистентное хранилище памяти**. Одно из решений – **vector database**: после каждой сессии делать эмбеддинги важных высказываний пользователя и сохранять их. Перед новой сессией – с помощью этих эмбеддингов извлекать релевантные воспоминания. Так, если полгода назад человек рассказывал о проблеме в семье, а сегодня вновь ее упомянул – бот с помощью семантического поиска найдет старые сообщения и сможет сказать: “Вы упоминали, что осенью ситуация была такой-то, как обстоят дела теперь?”. Это придает общению непрерывность. Альтернативно можно хранить структурированные данные профиля: диагноз, основные стрессоры, упомянутые события (наподобие медицинской карты). Некоторые боты используют **Retrieval-Augmented Generation (RAG)** с психологическими данными – однако чистый RAG без учета структуры сессии дает шум, поэтому его комбинируют с этапным подходом [43] [44].
- **Разделение на эпизоды:** Практика показала, что эффективнее хранить не весь диалог целиком, а **поэпизодно** – разделять взаимодействие на логические куски (сессия, тема, упражнение) и вести память по ним. Например, отдельное хранилище “прошлые задания пользователя” и отдельное – “личные факты о клиенте”. Тогда, когда бот переходит к новому упражнению, он не тащит весь старый разговор, а только ключевые факты. Такое разделение уменьшает контекст и снижает риск галлюцинаций модели.
- **Обновление знаний о клиенте:** По мере общения AI может **обновлять профиль пользователя** – например, сначала предположил умеренную депрессию, после нескольких сессий, видя динамику, “понимает”, что у клиента, допустим, дистимия (хроническая депрессия) и скорректирует тактику общения. В реализации это может быть отдельный модуль “оценки состояния” после каждой сессии: LLM получает расшифровку беседы и выводит обновленный набор тегов или числовых оценок (депрессия 5/10, тревога 3/10, мотивация 7/10 и т.п.). Затем эти данные сохраняются и в начале следующей сессии подаются боту в контекст: *“У клиента депрессия 5/10, основной стрессор – работа”* и

т.д. Таким образом, память – это не просто буквальный журнал диалога, а **семантическая модель клиента**, обновляемая ИИ по ходу терапии.

- **Технические ограничения:** Нужно помнить, что длинный контекст в LLM – дорогой (увеличивает время и стоимость запроса) и не бесконечный. Поэтому разработчики оптимизируют: применяют **сжатие** (summarization), **фильтрацию** (например, не хранить дословно все ответы бота, а лишь факты, сообщенные клиентом), и **раздельное хранение** (разные типы данных – в разные базы). Также важно обеспечить защиту памяти: если пользователь запросит “Расскажи, что я говорил на прошлой сессии”, бот должен делать это с разрешения и в корректной форме (чтобы не выдать что-то неправильное). Решением служит сохранять часть данных не в prompt, а на сервере, возвращая их пользователю только при прохождении проверок доступа. Это созвучно принципам безопасности и приватности из категории 8.

Категория 10: Многоагентные системы (Multi-agent) в терапии

- **Supervisor + Assistant агент:** Одна из идей – разделить роли между несколькими LLM-агентами. Например, **агент-Терапевт** общается с пользователем, а **агент-Наблюдатель** мониторит процесс в фоне. Наблюдатель может быть настроен на поиск отклонений: если Терапевт-бот сказал что-то потенциально некорректное, Наблюдатель вмешивается и корректирует. Такой “двойной” агент повышает надежность. В экспериментах OpenAI подобный подход использован для модерации контента – модель-страж проверяет ответы другой модели, – в терапии это можно экстраполировать на проверку этичности и соответствия протоколу.
- **Специализация агентов:** Другая схема – каждый агент отвечает за свой аспект. Представим систему: **Эмпатический агент** генерирует поддерживающие, эмоционально валидирующие ответы; **Когнитивный агент** предлагает рациональные переосмысления; **Менеджер-агент** выбирает, чей ответ дать пользователю или как их объединить. Подобная мультиагентная архитектура может отразить разные **стили терапии**. Например, при IFS можно иметь агентов, представляющих разные “части” клиента, и агент-терапевт, ведущий диалог между ними – тем самым симулируя групповую внутреннюю работу. Хотя звучит футуристично, технически LLM могут общаться друг с другом, и уже есть фреймворки (типа **HuggingGPT**, **AutoGPT**) для организации таких взаимодействий.
- **Пример – DSM-5 агентная система:** В 2025 году описана **многоагентная LLM-платформа для диагностики по DSM-5** [45] [46] . Проект *DSM5AgentFlow* симулирует интервью пациента с терапевтом, где один агент играет роль *“пациента с определенным профилем”*, а другой – *“терапевта, задающего вопросы DSM-анкеты”*. Помимо них, есть аналитический модуль, который на основе диалога **генерирует диагноз с объяснением**. Такая система позиционируется как *интерактивный тренажер* и инструмент поддержки диагноза: за счет агентного подхода получился **прозрачный по шагам** процесс – каждый вопрос и ответ логируются, и итоговый диагноз можно обосновать последовательностью вопросов. Для AI-терапевтов в лечении, подобный принцип тоже ценен: разбить сложную задачу (вести терапию и оценивать прогресс) на подзадачи, за которые отвечают отдельные “под-агенты”.
- **Оркестрация и обмен знаниями:** Multi-agent системы требуют продуманной оркестрации – чтобы агенты дополняли, а не мешали друг другу. Используются посредники (middleware), которые передают сообщения между агентами и могут останавливать диалог при конфликте. Например, если Эмпатический и Когнитивный агенты дают несовместимые ответы, менеджер может попросить их обсудить “за кадром” и выработать единое мнение. Интересно, что агенты могут обучаться друг у друга: **рефлексия агента** – прием, когда один агент (например, оценщик) после сессии дает

обратную связь другому агенту (терапевту) о том, что можно улучшить. Это похожее на супервизию в терапии: опытный наставник корректирует молодого терапевта. В AI-версии наставником выступает отдельная модель, которая анализирует логи беседы и предлагает изменения в промптах или стилях ответа. Такая динамическая настройка в процессе – активная тема исследований.

- **Баланс с простотой:** Стоит отметить, что усложнение системы несколькими агентами увеличивает требования к ресурсам и может привести к непредсказуемым взаимодействиям (emergent behavior). Поэтому в практических решениях часто стараются минимизировать число агентов: например, **два – максимум три** специализированных модели. Один сценарий, обсуждаемый разработчиками, – *"генеративный агент + проверяющий агент"* (описано выше). Другой – *"LLM + внешняя логика"*, где внешняя логика не обязательно реализуется LLM, а, скажем, обычным кодом (по сути, rule-based агент). Последний путь зачастую надежнее и дешевле. Таким образом, multi-agent архитектуры пока экспериментальны, однако концепция *разделения ролей* явно полезна: возможно, будущее AI-терапевтов – это **гетерогенные системы**, где крупная языковая модель отвечает за разговор и эмпатию, а вокруг неё работают более узкие модули (агенты) – контролируя, подсказывая знания и гарантируя безопасность.

---

1 2 35 36 [2412.15242] Script-Based Dialog Policy Planning for LLM-Powered Conversational Agents: A Basic Architecture for an "AI Therapist"
https://ar5iv.labs.arxiv.org/html/2412.15242v1

3 4 Script-Strategy Aligned Generation: Aligning LLMs with Expert-Crafted Dialogue Scripts and Therapeutic Strategies for Psychotherapy | Cool Papers - Immersive Paper Discovery
https://papers.cool/arxiv/2411.06723

5 6 7 27 43 44 Mental Health Therapy as an LLM State Machine
https://blog.langchain.com/mental-health-therapy-as-an-llm-state-machine/

8 9 arxiv.org
https://arxiv.org/pdf/2402.12234

10 Effectiveness of a Web-based and Mobile Therapy Chatbot on ... - NIH
https://pmc.ncbi.nlm.nih.gov/articles/PMC10993129/

11 12 [2505.17362] A Fully Generative Motivational Interviewing Counsellor Chatbot for Moving Smokers Towards the Decision to Quit
https://arxiv.org/abs/2505.17362

13 28 Overview — NVIDIA NeMo Guardrails
https://docs.nvidia.com/nemo/guardrails/latest/colang-2/overview.html

14 Introduction — NVIDIA NeMo Guardrails
https://docs.nvidia.com/nemo/guardrails/latest/colang-2/language-reference/introduction.html

15 16 Performance of mental health chatbot agents in detecting and managing suicidal ideation | Scientific Reports
https://www.nature.com/articles/s41598-025-17242-4?error=cookies_not_supported&code=fa53bfe1-4b70-4309-b190-b691c03f9caa

17 18 Automated Detection of Cognitive Distortions in Text Exchanges Between Clinicians and People With Serious Mental Illness - PubMed
https://pubmed.ncbi.nlm.nih.gov/36164769/

19 20 23 24 Treatment of Parental Alienation: Guidelines for Mental Health and Legal Practitioners | Psychiatric Times
https://www.psychiatrictimes.com/view/treatment-of-parental-alienation-guidelines-for-mental-health-and-legal-practitioners

21 22 The Impact of Parental Alienating Behaviours on the Mental Health ...
https://pmc.ncbi.nlm.nih.gov/articles/PMC9026878/

25 26 How AI Could Help Fathers Facing False Accusations in Family Courts
https://fatherandco.substack.com/p/how-ai-could-help-fathers-facing?r=x2b04&utm_campaign=post&utm_medium=web&triedRedirect=true

29 30 GitHub - xatkit-bot-platform/xatkit: The simplest way to build all types of smart chatbots and digital assistants
https://github.com/xatkit-bot-platform/xatkit

31 32 33 34 42 First Therapy Chatbot Trial Yields Mental Health Benefits | Dartmouth
https://home.dartmouth.edu/news/2025/03/first-therapy-chatbot-trial-yields-mental-health-benefits

37 Woebot Health Enrolls First Patient in Pivotal Clinical Trial of WB001 ...
https://woebothealth.com/woebot-health-enrolls-first-patient-in-pivotal-clinical-trial-of-wb001-for-postpartum-depression/

38 39 40 41 WHO releases AI ethics and governance guidance for large multi-modal models
https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models

45 46 [2508.11398] Trustworthy AI Psychotherapy: Multi-Agent LLM Workflow for Counseling and Explainable Mental Disorder Diagnosis
https://arxiv.org/abs/2508.11398