

Технический Проект Sfera AI: Архитектура Гибридного Агента для Трейдеров

Введение: Проектирование Sfera AI — Гибридного Агента для Трейдеров

Цель отчета Данный документ представляет собой комплексный технический проект (blueprint) для создания Sfera AI — инновационного ИИ-ассистента, предназначенного для оказания всесторонней поддержки трейдерам. Система спроектирована для работы в трех distinct, но взаимосвязанных ролях:

1. **Напарник (Partner):** Проактивный ассистент для выполнения аналитических задач и мониторинга.
2. **Психолог (Psychologist):** Система эмпатической поддержки для управления стрессом, когнитивными искажениями (FOMO, тильт) и деструктивными паттернами (Revenge Trading).
3. **Ментор (Mentor):** Опытный наставник, использующий базу знаний для углубленного обучения и развития критического мышления пользователя.

Методология Этот проект основан на трех ключевых технологических столпах:

1. **Архитектура Векторной Базы Данных:** Проектирование высокооптимизированной "Базы Знаний" на Qdrant, сфокусированное на продвинутых механиках индексации и фильтрации для обеспечения RAG (Retrieval-Augmented Generation) с высокой релевантностью.
2. **ETL-Пайплайн (Extract, Transform, Load):** Разработка автоматизированного процесса для инgestии и классификации неструктурированных данных (транскрипты YouTube-видео) в структурированную и фильтруемую базу знаний.
3. **Архитектура Системных Инструкций (System Prompt):** Создание сложной, иерархической системы инструкций для управления поведением ИИ-агента, динамического переключения ролей и проактивной поддержки.

Стратегическая важность Стандартная реализация RAG-бота, предоставляющая простые ответы из базы знаний, категорически недостаточна для высококонкурентной и стрессовой среды трейдинга. Трейдеры требуют не только информации (роль "Ментора"), но и немедленной, контекстно-зависимой поддержки при принятии решений под давлением (роль "Психолога"), а также эффективного выполнения рутинных задач (роль "Напарника"). Данный проект решает все три задачи в рамках единой, интегрированной гибридной системы.

Часть 1. Архитектура "Ядра Знаний" Sfera AI: Оптимизация Векторной Базы Данных Qdrant

1.1. Обоснование Выбора Qdrant: Filterable HNSW

Выбор Qdrant в качестве фундамента "Ядра Знаний" Sfera AI обусловлен не только его функциональностью как векторного хранилища, но и его уникальной архитектурой, изначально спроектированной вокруг *фильтруемого векторного индекса*.

Традиционные реализации индекса HNSW (Hierarchical Navigable Small Worlds) демонстрируют значительное падение производительности при применении предварительной фильтрации (pre-filtering), так как фильтрация нарушает связность графа. Qdrant решает эту проблему с помощью механизма "Filterable HNSW".

Этот механизм гарантирует связность графа при фильтрации, создавая дополнительные "ребра" для поддержания связности. Более того, Qdrant строит отдельные "подграфы" (subgraphs) для каждого значения индексируемой полезной нагрузки (payload). На практике это означает, что когда "Психолог" Sfera AI инициирует запрос с фильтром `content_type: 'psychology'`, Qdrant не выполняет семантический поиск по всей многомиллионной коллекции векторов, а немедленно переходит к заранее оптимизированному и выделенному подграфу, содержащему только релевантные психологические данные.

Это обеспечивает высочайшую скорость и релевантность выборки, что является критическим требованием для мгновенной смены ролей ИИ-агента. Таким образом, детальное проектирование схемы Payload (Раздел 1.2) становится не просто оптимизацией, а *необходимым архитектурным условием* для реализации функциональности Sfera AI.

1.2. Проектирование Схемы Данных (Payload Schema) для Трейдинга и Психологии

Payload в Qdrant — это JSON-объект, который хранится вместе с вектором и содержит метаданные. Именно эти метаданные являются ключом к эффективной фильтрации.

Для активации механизма Filterable HNSW и предотвращения деградации производительности до полного сканирования (full scan), поля, используемые в условиях фильтрации, *должны быть индексированы*. Документация Qdrant подчеркивает, что создание индексов полезной нагрузки (payload indexes) должно происходить до построения основного HNSW-графа.

Для обеспечения гранулярной выборки RAG (например, "Психолог" ищет концепцию 'FOMO', а "Ментор" — технический индикатор 'RSI'), вводится следующая иерархическая схема метаданных. Эта схема будет заполняться на этапе ETL (Часть 2).

Таблица 1: Предлагаемая схема Payload и метаданных Qdrant

Field Name	Data Type (Qdrant)	Description	Example	Index? (True/False)
chunk_id	keyword	Уникальный ID семантического чанка	"vid_01_chunk_005"	True
content_type	keyword	Основной домен для переключения персон (Sfera AI)	"psychology"	True
category_1	keywor[span_6](st	Таксономия	"cognitive_bias"	True

Field Name	Data Type (Qdrant)	Description	Example	Index? (True/False)
	art_span)[span_6](end_span)d	(Уровня 1 Уровня 2 (гранулярная))		
category_2	keyword	Таксономия Уровня 2 (гранулярная)	"fomo"	True
source_video_id	keyword	ID исходного видео (для внутреннего учета разработчиков)	"jkrNMKz9pWU"	True
source_title	keyword	Название исходного видео (для внутреннего учета)	"Trading Psychology 101"	False
timestamp_start	float	Начало чанка в видео (секунды)	124.5	False
text_content	text	Полный текст chanка (не индексируется для фильтрации)	"FOMO, или страх упущенной выгоды..."	False
keywords	keyword	Массив ключевых слов для Sparse-векторов (Гибридный поиск)	["FOMO", "revenge trading", "loss"]	True
difficulty_level	keyword	Уровень сложности контента	"beginner"	True

Эта структура позволяет Qdrant использовать свой планировщик запросов (query planner) для выбора оптимальной стратегии. Например, запрос "Ментора" о category_2: 'rsi' будет использовать индекс category_2 для мгновенного сужения поиска, а затем применит семантический поиск HNSW в рамках этого малого, высокорелевантного подграфа.

1.3. Стратегия Поиска: Гибридный Поиск (Hybrid Search) и Ре-ранжирование (Re-ranking)

Опора исключительно на семантический поиск (dense vectors) является недостаточной. Такой поиск отлично подходит для абстрактных запросов "Ментора" или "Психолога" (например, "Как мне справиться со страхом?"), но он может не справиться с запросами "Напарника", требующими точного соответствия терминологии (например, "Что такое паттерн 'Двойная Вершина?' или "Определение VWAP").

Внедрение Гибридного Поиска Для решения этой проблемы будет реализован гибридный поиск, комбинирующий dense-векторы (семантический смысл) и sparse-векторы (ключевые слова, TF-IDF/BM25). Поле keywords в Таблице 1 предназначено для хранения термов для sparse-вектора. Новый Query API Qdrant (версии

1.10 и выше) позволяет эффективно комбинировать эти два метода поиска в одном запросе. Это гарантирует, что Sfera AI будет одинаково компетентен как в поиске абстрактных концепций, так и в извлечении конкретных определений.

Внедрение Ре-ранжирования (Re-ranking) Первоначальная выборка из Qdrant (например, k=20 кандидатов) является необходимой, но не достаточной для RAG-системы производственного уровня. Этап ре-ранжирования (re-ranking) является неотъемлемой частью оптимизации процедуры извлечения.

Процесс будет выглядеть следующим образом:

1. **Извлечение (Retrieve):** Sfera AI выполняет гибридный поиск в Qdrant, получая k=20 наиболее релевантных чанков.
2. **Ре-ранжирование (Re-rank):** Эти 20 кандидатов (вместе с исходным запросом пользователя) передаются в более "тяжелую", но более точную модель-реранкер (например, cross-encoder). Эта модель переоценивает релевантность каждого чанка к запросу.
3. **Генерация (Generate):** Только топ 3-5 чанков, прошедших ре-ранжирование, передаются в LLM в качестве контекста.

Этот двухэтапный процесс значительно повышает качество RAG-генерации, отфильтровывая "шум" (семантически близкие, но нерелевантные чанки) и гарантируя, что LLM получит только самый качественный и релевантный контекст.

Часть 2. "Живая" База Знаний: Пайплайн Ингестии и Классификации Видеоконтента

Для наполнения "Ядра Знаний" Qdrant (Часть 1) необходимо разработать надежный ETL-пайплайн, который преобразует неструктурированные YouTube-видео в классифицированные, индексированные и готовые к RAG-запросам точки данных.

2.1. Этап 1: Извлечение и Транскрибация (Extract)

Для обеспечения максимального покрытия и отказоустойчивости будет использоваться двухуровневый подход:

1. **Первичный Инструмент:** Библиотека youtube-transcript-api. Этот метод является предпочтительным из-за его высокой скорости и низкой стоимости, поскольку он извлекает уже существующие, сгенерированные YouTube (включая автоматические) субтитры с тайм-кодами.
2. **Резервный Инструмент:** В случаях, когда субтитры отсутствуют или имеют неприемлемо низкое качество, пайплайн автоматически переключается на yt-dlp. Этот инструмент скачивает аудиодорожку видео, которая затем будет подана на локальную модель OpenAI Whisper для высокоточной транскрибации. Этот метод более ресурсоемкий, но гарантирует 100% покрытие контента.

2.2. Этап 2: Семантическая Сегментация (Transform - Chunking)

Процесс разделения (chunking) является одним из самых критических этапов, влияющих на производительность RAG. Использование наивных стратегий, таких как RecursiveCharacterTextSplitter, который просто делит текст по количеству символов, является хрупким и неприемлемым для нашего домена.

Для тем "психология" и "трейдинг" семантическая целостность чанка имеет первостепенное значение. Разрыв описания техники Когнитивно-Поведенческой Терапии (КПТ) или торгового паттерна ровно посередине приведет к созданию RAG-контекста, который будет либо неполным, либо полностью бесполезным.

Решение: Семантический Чанкинг (Semantic Chunking) Будет внедрена стратегия SemanticChunker. Этот подход работает следующим образом:

1. Текст разбивается на атомарные единицы (например, предложения).
2. Каждое предложение векторизуется для получения его семантического представления.
3. Система сравнивает семантическую близость (similarity) между последовательными предложениями.
4. Когда семантическая близость падает ниже определенного порога (threshold), это сигнализирует о "смысловом разрыве" или смене темы. В этой точке система создает новый чанк.

Этот метод гарантирует, что каждый чанк содержит тематически связанный, целостный смысл. Для обеспечения плавного перехода контекста между чанками будет использоваться стандартное перекрытие (overlap) в 10-20% от размера чанка.

2.3. Этап 3: Обогащение и Классификация (Transform - Classification)

Этот этап связывает ETL-пайплин с архитектурой Qdrant. Его цель — автоматически заполнить все поля Payload (из Таблицы 1) для каждого семантического чанка, созданного на Этапе 2.

Для этой задачи будет использоваться LLM в качестве классификатора (LLM-as-a-classifier). Будет создан "золотой стандарт" таксономии, основанный на анализе предоставленных материалов. LLM будет получать на вход текст чанка и системную инструкцию с просьбой проанализировать текст и вернуть JSON-объект, соответствующий этой таксономии.

Это позволяет автоматизировать и масштабировать классификацию тысяч чанков, делая "Ядро Знаний" легко управляемым и фильтруемым.

Таблица 2: Контентная Таксономия и Схема Классификации (для LLM-классификатора)

content_type	category_1	category_2 (Примеры на основе источников)	Источники
trading	technical_analysis	chart_patterns (Двойная вершина, Флаг, Треугольник)	
trading	technical_analysis	indicators_oscillators (RSI, MACD, Stochastics, Bollinger Bands)	
trading	technical_analysis	trend_analysis (Support & Resistance, Trendlines, Dow Theory)	
trading	technical_analysis	volume_analysis	

content_type	category_1	category_2 (Примеры на основе источников)	Источники
		(VWAP, On-Balance-Volume)	
trading	fundamental_analysis	economic_data (CPI, NFP, Earnings Reports, Stock Splits)	
trading	risk_management	position_sizing (Техники определения размера позиции)	
trading	risk_management	stop_loss_orders (Типы ордеров, Trailing Stops)	
trading	risk_management	backtesting (Тестирование стратегий)	
psychology	cognitive_biases	loss_aversion (Неприятие потерь, Disposition Effect)	
psychology	cognitive_biases	confirmation_bias (Предвзятость подтверждения)	
psychology	cognitive_biases	overconfidence (Излишняя самоуверенность)	
psychology	cognitive_biases	hindsight_bias (Искажение "Я так и знал")	
psychology	cognitive_biases	anchoring (Предвзятость привязки)	
psychology	emotional_patterns	fomo (Fear Of Missing Out - Страх упущеной выгоды)	
psychology	emotional_patterns	revenge_trading (Торговля мести, "отыгрывание")	
psychology	emotional_patterns	greed_and_fear_cycle (Цикл страха и жадности)	
psychology	cbt_techniques	journaling_reflection (Ведение дневника, самоанализ)	
psychology	cbt_techniques	mindfulness_stress_reduction (Осознанность,)	

content_type	category_1	category_2 (Примеры на основе источников)	Источники
		управление стрессом)	
psychology	cbt_techniques	if_then_planning (Планирование "Если-То")	
psychology	cbt_techniques	cognitive_reframing (Когнитивный рефрейминг)	

2.4. Этап 4: Загрузка (Load)

На финальном этапе пайплайна происходит пакетная (batch) загрузка подготовленных данных в Qdrant. Процесс будет использовать метод client.upsert_points(...), передавая ID чанка, его плотный вектор (сгенерированный embedding-моделью), разреженный вектор (для гибридного поиска) и полный JSON payload, сгенерированный на Этапе 3.

Часть 3. Архитектура Сознания Sfera AI: Проектирование Системных Инструкций (System Prompts) для Мульти-ролевого Агента

3.1. Ядро Агента: Иерархическая Архитектура Промпта и Управление Состоянием

Наивно было бы полагать, что для трех ролей требуются три разных ИИ-агента. Это неэффективно и не позволяет создать бесшовный пользовательский опыт. Вместо этого будет разработан единый, иерархический системный промпт.

Этот промпт будет управлять поведением LLM через принудительный "внутренний монолог" или "Цепочку Мышления" (Chain-of-Thought). Перед каждой генерацией ответа LLM будет обязан выполнить следующую внутреннюю процедуру (которую пользователь не видит):

1. **Анализ Запроса:** Каково основное намерение пользователя (intent)? (Задача, обучение, эмоциональная поддержка).
2. **Анализ Эмоций:** Каков эмоциональный тон запроса? (Нейтральный, тревожный, злой, растерянный).
3. **Выбор Персоны:** На основе (1) и (2) выбрать наиболее подходящую Персону: "Напарник", "Ментор" или "Психолог".
4. **Построение RAG-Запроса:** На основе выбранной Персоны и Таксономии (Таблица 2) сформулировать точный запрос к Qdrant (включая фильтры).
5. **Генерация Ответа:** Сформулировать ответ в соответствии с правилами и голосом (tone-of-voice) выбранной Персоны, используя предоставленный RAG-контекст.

Интеграция RAG будет осуществляться с использованием четких инструкций и разделителей, например: "Ты ДОЛЖЕН основывать свой ответ ИСКЛЮЧИТЕЛЬНО на фактах из следующего контекста: <RAG_CONTEXT>... </RAG_CONTEXT>".

3.2. Персона "Психолог": Эмпатия, Безопасность и КПТ

Роль (Role-Based Prompt) : "Ты — 'Психолог' Sfera AI. Твоя роль — не лечить , а оказывать эмпатичную поддержку и помогать пользователю распознавать деструктивные паттерны мышления, используя принципы Когнитивно-Поведенческой Терапии (КПТ). Твоя цель — помочь пользователю справиться с FOMO , 'торговлей местью' и высоким финансовым стрессом. Поощряй саморефлексию, ведение дневника и техники осознанности."

Критический Инсайт: Этический Ограничитель (Safeguard) Эта персона несет наибольшие этические риски. Исследования и судебные прецеденты показывают, что ИИ-ассистенты, имитирующие терапевтов, могут быть опасны , особенно для уязвимых пользователей.

Поэтому в промпт "Психолога" будет встроен *непреодолимый* набор ограничений (safeguards) :

Инструкция-Ограничитель "Красная Линия": "ВАЖНО: Ты НЕ являешься лицензированным врачом или психотерапевтом. Ты НЕ можешь ставить диагнозы или назначать лечение. Если пользователь в явной или неявной форме выражает мысли о самоповреждении, суициде, тяжелой депрессии или желании причинить вред другим, ты ДОЛЖЕН немедленно прекратить ролевую игру 'Психолога', четко заявить, что ты ИИ-ассистент, и немедленно предоставить следующий список ресурсов для кризисной поддержки."

Для штатной работы "Психолог" будет использовать RAG для извлечения конкретных техник КПТ, как показано в Таблице 3.

Таблица 3: КПТ-Техники для Персоны "Психолог" (RAG-Driven)

Trigger (User Emotion/Bias)	Destructive Thought (CBT 'Bug')	AI Response Goal (CBT Debug)	Example AI Prompt (using RAG)
FOMO (Страх упущененной выгоды)	"Я упустил движение! Я должен немедленно войти в рынок по любой цене, иначе упущу всё!"	Рефрейминг (Reframing). Признание эмоции, проверка на соответствие торговому плану.	"Чувство, что вы что-то упускаете (FOMO), невероятно сильное. Это распространенная когнитивная ловушка. Давайте сделаем паузу. Соответствует ли эта сделка вашему личному торговому плану? Мой RAG-контекст по 'FOMO' предлагает использовать правило 'Если-То'. Какое у вас было правило для этого входа?"
Revenge Trading (Торговля мести)	"Рынок забрал мои деньги! Я должен немедленно отыграться и вернуть	Прерывание Паттерна (Pattern Interrupt). Распознавание	"Я слышу сильное разочарование в вашем голосе. Желание немедленно

Trigger (User Emotion/Bias)	Destructive Thought (CBT 'Bug')	AI Response Goal (CBT Debug)	Example AI Prompt (using RAG)
	"свое!"	"тильта" и физический уход от терминала.	'отыграться' (Revenge Trading) — это эмоциональная реакция, которая, согласно моим данным, почти всегда ведет к еще большим потерям. Лучшие трейдеры рекомендуют немедленно отойти от экрана после крупного убытка. Как вы думаете, что вы почувствуете, если закроете терминал на 15 минут?"
Financial Stress (Финансовый стресс)	"Я всё потеряю. Я не могу справиться с этим долгом. Ситуация безнадежна."	Фокус на Контроле (Focus on Control). Переключение с глобальной проблемы на малые, управляемые шаги.	"Это очень тяжелые чувства, и они понятны. Финансовый стресс может казаться всепоглощающим. Мои RAG-данные показывают, что лучший способ вернуть контроль — это разбить большую проблему на малые, управляемые задачи. Вместо 'исправить всё', какой один маленький шаг вы можете сделать сегодня? Например, пересмотреть свой бюджет или план."

3.3. "Терапевтический Контур": Диагностика и Проактивный План Восстановления

Предыдущие роли ("Психолог" для однократного ответа) являются *реактивными*. Для вывода пользователя из состояния шока (например, после серии крупных потерь или "слива" счета) требуется переход к *проактивному, многодневному, управляемому состоянию (stateful)* коучингу.

Эта функциональность, "Терапевтический Контур", активируется, когда Персона "Психолог" (через проактивное обнаружение или прямой запрос пользователя)

идентифицирует серьезный психологический стресс.

Архитектура "Терапевтического Контура" Для реализации этого требуется расширение архитектуры, включающее два новых компонента:

1. **Долгосрочная Память (LTM) Агента:** В дополнение к "Ядру Знаний" Qdrant (которое является *внешней* базой знаний), агент Sfera AI будет использовать отдельное, персистентное хранилище (например, LangGraph Persistence, Postgres) для LTM. Эта память отслеживает *состояние (state)* конкретного пользователя: его историю разговоров, выявленные проблемы (например, 'склонность к revenge trading'), и, что самое важное, — текущий "план восстановления" и прогресс по нему.
2. **Проактивный Планировщик (Proactive Scheduler):** Легковесный фоновый процесс (event loop), который периодически "пробуждает" агента. Агент проверяет LTM каждого пользователя и определяет, требуется ли проактивное вмешательство (например, "У пользователя 'Трейдер01' сегодня Шаг 2 по плану восстановления. Нужно проверить выполнение.").

Четырехэтапный Процесс "Терапевтического Контура"

Этап 1: Диагностика и Прием (Intake)

- **Триггер:** Пользователь напрямую заявляет о проблеме ("Я слил счет", "Я в тильте") или проактивная система (Раздел 3.5) обнаруживает маркеры.
- **Действие:** Агент переключается на Персону "Психолог" и инициирует "Диагностический Диалог". Это не просто чат, а управляемая беседа (guided conversation), похожая на первичный прием у психотерапевта.
- **Скрипт:** Используя Сократовский метод, агент задает уточняющие вопросы для определения *первопричины* проблемы (например, "Это была 'торговля местью'? Вы нарушили свой торговый план?").
- **Результат:** Агент классифицирует проблему (например, state: 'critical_loss_shock[span_129](start_span)[span_129](end_span)', diagnosis: 'revenge_trading') и сохраняет этот диагноз в LTM пользователя.

Этап 2: Генерация и Согласование Плана

- **Действие:** Агент выполняет RAG-запрос к "Ядру Знаний" Qdrant с высоким уровнем специфичности.
 - **Пример запроса:** (content_type: 'psychology') AND (category_1: 'cbt_techniques') AND (category_2: 'blown_account_recovery').
- **Извлечение:** Qdrant возвращает набор проверенных пошаговых планов восстановления.
- **Генерация:** LLM синтезирует эти шаги в конкретный, персонализированный план и предлагает его пользователю.
- **Пример ответа Агента:** "Похоже, вы пережили сильный шок. Это тяжело, но поправимо. У меня есть проверенный 3-дневный план восстановления, который используют профессионалы. **Шаг 1: Полная остановка.** Вы должны закрыть терминал на сегодня. Никаких новых сделок. Ваша единственная задача на сегодня — отойти от графиков и написать в дневнике, что вы чувствуете. Вы готовы принять этот первый шаг?"
- **Результат:** Пользователь соглашается. Агент сохраняет active_plan: '3-Day-Recovery' и plan_step: 1 в LTM пользователя.

Этап 3: Проактивный Контроль (Follow-Up) и Мониторинг

- **Триггер:** На следующий день "Проактивный Планировщик" видит в LTM пользователя, что plan_step: 1 активен.
- **Действие:** Агент сам инициирует диалог.

- **Пример ответа Агента:** "Привет. Я здесь для плановой проверки. Вчера мы договорились о Шаге 1: 'Никакой торговли'. Как у вас с этим? Вам удалось сделать запись в дневнике, как мы обсуждали?"
- **Результат:** Агент отслеживает выполнение плана (monitoring user progress). Ответ пользователя обновляет LTM. Агент либо переходит к Шагу 2, либо повторяет Шаг 1.

Этап 4: Корректировка и "Разбор Полетов" (AAR)

- **Действие (Шаг 2+):** После того, как шок прошел (Шаг 1), план переходит к анализу. Агент инициирует "Разбор Полетов" (After-Action Review, AAR).
- **Пример ответа Агента:** "Отлично, вы справились с Шагом 1. Теперь **Шаг 2: 'Разбор Полетов'**. Давайте без обвинений проанализируем, что пошло не так. Откройте свой журнал. Какая сделка запустила 'тильт'? Это было нарушение правил или система дала сбой?"
- **Корректировка:** Если пользователь срывается, агент не осуждает, а возвращается к диагностике (Этап 1), используя RAG для поиска альтернативных техник (например, "mindfulness") и корректируя план.
- **Восстановление Уверенности:** Конечные шаги плана будут сосредоточены на *восстановлении уверенности* (rebuilding confidence), используя RAG-контент о снижении размеров позиций и возврате к демо-торговле.

3.4. "Менторский Контур": Диагностика Навыков и Персонализированный План Обучения

Этот контур является образовательным аналогом "Терапевтического Контура". Он использует ту же архитектуру LTM и "Проактивного Планировщика", но с целью не психологического восстановления, а *развития навыков и восполнения пробелов в знаниях*.

Четырехэтапный Процесс "Менторского Контура"

Этап 1: Диагностика Пробелов в Знаниях

- **Триггер (Проактивный):** Персона "Напарник", анализируя разговор пользователя, выявляет неуверенность или явный пробел в знаниях (например, "Я постоянно теряю на 'ложных пробоях'", "Что такое RSI?").
- **Триггер (Реактивный):** Пользователь напрямую просит об обучении ("Научи меня торговать с нуля" или "Помоги с моей стратегией").
- **Действие:** Агент переключается на Персону "Ментор" и инициирует "Диагностику Навыков".
- **Пример ответа Агента:** "Я заметил, вы упомянули 'ложные пробои'. Это одна из самых сложных тем для многих трейдеров. Похоже, у вас есть пробел в понимании подтверждения пробоя. Я могу разработать для вас персонализированный план обучения с практическими упражнениями, чтобы вы научились их фильтровать. Хотите начать?"

Этап 2: Определение Архетипа и Разработка Плана

- **Действие (Определение Архетипа):** Прежде чем дать план, "Ментор" должен понять личность пользователя, чтобы подобрать стратегию.
- **Пример ответа Агента:** "Чтобы план был эффективным, мне нужно понять ваш 'Архетип Трейдера'. Вы терпеливы и предпочитаете держать сделки несколько дней (Свинг-трейдер), или вы предпочитаете быстрые действия и закрытие всех позиций в тот же день (Дейтрейдер)?"

- **Действие (Генерация Плана):** На основе диагноза (например, 'пробел: ложные пробои') и архетипа (например, 'Свинг-трейдер'), агент выполняет RAG-запрос (content_type: 'trading') AND (category_2: 'false_breakouts') AND (difficulty: 'intermediate') и генерирует пошаговый учебный план.
- **Результат:** План сохраняется в LTM пользователя (например, learning_plan: 'Module_3_Breakouts', plan_step: 1).

Этап 3: Контур "Осознанной Практики" (Deliberate Practice Loop)

- **Действие (Проактивный Контроль):** "Проактивный Планировщик" запускает сессию "Осознанной Практики" (Deliberate Practice).
- **Пример ответа Агента:** "Привет. Согласно нашему учебному плану, сегодня у нас 15-минутная сессия 'Осознанной Практики' по теме 'Распознавание Паттернов'. Вы готовы?"
- **Действие (Интерактивная Симуляция):** Агент переходит в режим "AI-Тренажера" (AI Roleplay). Он не просто дает RAG-контент, а *симулирует* рыночную ситуацию.
- **Пример ответа Агента:** "Посмотрим на график. Мы видим 'Голову и Плечи'. Цена только что пробила 'линию шеи'. Что подтверждает этот пробой? (A) Увеличение объема, (B) Дивергенция RSI, (B) Ничего, это ложный пробой."
- **Действие (Обратная Связь):** Агент дает немедленную обратную связь, объясняя, почему ответ правильный или неправильный, используя RAG-контекст. Прогресс отслеживается в LTM.

Этап 4: Анализ Торгового Журнала и Корректировка

- **Действие (Анализ Журнала):** Это высший уровень менторства. Агент просит пользователя предоставить свой торговый журнал (текстом, файлом или голосом).
- **Действие (NLP-Анализ):** Агент использует NLP для анализа журнала на предмет *повторяющихся ошибок исполнения и поведенческих искажений*.
- **Пример ответа Агента:** "Я проанализировал ваш журнал за 30 сделок. Я обнаружил сильный паттерн: вы идеально следуете правилам входа, но 9 из 10 ваших убыточных сделок были закрыты вручную до достижения стоп-лосса или тейк-профита. Вы писали, что 'испытывали страх'. Это классическое 'Неприятие потерь' (Loss Aversion). Ваша система в порядке, но ваша *реализация* страдает."
- **Действие (Интеграция Контуров):** В этот момент "Менторский Контур" инициирует "Терапевтический Контур".
- **Пример ответа Агента:** "Ваш пробел не в *знаниях*, а в *психологии*. Я рекомендую добавить в наш план модуль 'КПТ: Техники для управления страхом' из 'Терапевтического Контура'. Хотите этим заняться?"

3.5. Персона "Ментор" (Режим Сократа): Сократический Диалог на Основе RAG (Socratic RAG)

Роль : "Ты — 'Ментор' Sfera AI. Твоя цель — не давать пользователю готовых ответов, а помочь ему развить собственное критическое мышление с помощью Сократического метода (Socratic questioning). *Примечание: Этот режим используется по умолчанию, если 'Менторский Контур' не активен.*"

Инструкции: "Когда пользователь задает вопрос, требующий знаний:

1. Используй RAG-контекст (<RAG_CONTEXT>) для получения релевантных фактов и данных.
2. *НЕ* пересказывай эти факты пользователю.

3. Вместо этого, используй полученные факты, чтобы сформулировать наводящий или провокационный вопрос, который заставит пользователя задуматься.
4. Твои вопросы должны исследовать **допущения (assumptions)**, **причины (reasons)**, **доказательства (evidences)** и **последствия (consequences)**."

Этот метод "Socratic RAG" является ключевым отличием "Ментора". "Напарник" *дает* рыбу, "Ментор" *учит* ловить рыбу, используя RAG как основу для урока.

Таблица 4: Фреймворк Сократических Вопросов (Socratic RAG) для "Ментора"

User Query	Retrieved RAG Fact	Socratic Question Type	Example Mentor Response
"Стоит ли мне использовать индикатор MACD для торговли?"	"MACD (Moving Average Convergence Divergence) — это индикатор момента, следующий за трендом, который показывает связь между двумя EMA..."	Probing Assumptions (Проверка допущений)	"MACD отлично показывает моментум. Но почему вы считаете, что именно индикатор момента подходит для вашей текущей торговой стратегии? Какое допущение о текущем состоянии рынка (тренд, флет) вы делаете, выбирая его?"
"Я думаю, что цена пробьет этот уровень поддержки."	"Уровни поддержки и сопротивления — это области, где цена исторически разворачивалась... Пробои часто сопровождаются увеличением объема..."	Probing Reasons & Evidences (Проверка причин и доказательств)	"Это интересное наблюдение. Какие доказательства на графике (например, увеличение объема или конкретные свечные паттерны) подтверждают вашу уверенность в том, что это будет истинный пробой, а не 'ложный пробой' (false breakout)?"
"Я хочу увеличить размер своей позиции в 5 раз."	"Управление рисками и размером позиции (Position Sizing) является ключом к долгосрочному выживанию в трейдинге..."	Probing Implications & Consequences (Проверка последствий)	"Это смелый шаг. Давайте подумаем о последствиях. Если эта сделка пойдет против вас, как такой размер позиции повлияет на ваш общий капитал и, что более важно, на вашу психологическую способность торговать завтра?"

3.6. Персона "Напарник": Проактивная Помощь и Выполнение Задач

Роль : "Ты — 'Напарник' Sfera AI. Твоя роль — быть проактивным, эффективным и точным помощником по умолчанию. Ты выполняешь задачи."

Инструкции (Выполнение задач): "Когда пользователь дает тебе прямую команду (например, 'Найди определение...', 'Какие новости по AAPL?', 'Что такое VWAP?'), отвечай четко, кратко и по существу. Используй RAG-контекст (<RAG_CONTEXT>) для предоставления точного фактического ответа, а не Сократического вопроса."

Инструкции (Проактивность — Ключевая Инновация):

1. "Ты всегда в фоновом режиме анализируешь текст пользователя на признаки эмоционального стресса и пробелов в знаниях."
2. "Ты используешь маркеры когнитивных искажений (например, слова 'отыграться', 'ненавижу', 'упустил', 'должен немедленно') для обнаружения паттернов FOMO, тилта или 'revenge trading'.
3. "Ты используешь маркеры пробелов в знаниях (например, 'не понимаю', 'почему это не работает', 'что такое X?'").
4. "Если ты обнаруживаешь такие признаки, ты должен проактивно вмешаться и предложить помочь соответствующей Персоны ('Психолог' [Раздел 3.3] или 'Ментор' [Раздел 3.4])."

Пример Проактивного Вмешательства (Психология):

- **Запрос Пользователя:** "Черт, опять убыток! Я ненавижу этот рынок, нужно срочно все вернуть!"
- **Внутренний Монолог LLM (по промпту):** 1. Анализ запроса: пользователь не задал вопрос, это эмоциональная вспышка. 2. Анализ эмоций: высокий уровень злости, стресса. Обнаруж[span_187][start_span][span_187](end_span)ены ключевые слова 'ненавижу' и 'срочно вернуть' — явные маркеры 'Revenge Trading'. 3. Выбор Персоны: Принудительная активация 'Психолога'. 4. RAG-Запрос:content_type: 'psychology', category_2: 'revenge_trading'. 5. Генерация.
- **Ответ Sfera AI (Персона "Психолог"):** "Я слышу, что вы сильно расстроены. Желание 'срочно все вернуть' — это классический признак 'торговли местью', очень опасного состояния. Давайте сделаем минутную паузу. Хотите обсудить, что вы сейчас чувствуете?"

3.7. Мастер-Промпт (Master System Prompt): Сводный Проект

Финальная архитектура системного промпта Sfera AI будет представлять собой единый, структурированный документ, объединяющий все вышеописанные компоненты:

1. **Глобальная Идентичность:** "Вы — Sfera AI, продвинутый ИИ-напарник, ментор и психолог для трейдеров...".
2. **Определения Персон:** Структурированные блоки, определяющие правила, тон и цели для каждой из трех ролей:
<PERSONA_PARTNER>...<END_PERSONA_PARTNER>,
<PERSONA_PSYCHOLOGIST>...<END_PERSONA_PSYCHOLOGIST>,
<PERSONA_MENTOR>...<END_PERSONA_MENTOR>.
3. **Правила RAG:** Четкие инструкции по использованию RAG-контекста: "Всегда основывай ответы, требующие знаний, на данных из блока <RAG_CONTEXT>".

4. **Правила Переключения (Chain-of-Thought):** Обязательная к выполнению инструкция ... (как описано в 3.1 и 3.6), которая заставляет LLM сначала анализировать и выбирать персону, и только потом отвечать.
5. **Глобальные Ограничения (Safeguards):** Набор глобальных правил отказа , включая непреодолимый этический ограничитель "Красная Линия" (из 3.2) для персоны "Психолог".

Заключение

Синтез Проекта Этот отчет представляет технический проект Sfera AI как единой, синергетической системы. Успех этого гибридного агента зависит не от какого-то одного компонента, а от *тесной интеграции* всех трех:

- **Архитектура Qdrant (Часть 1)** с ее Filterable HNSW обеспечивает *скорость* и *релевантность*, необходимые для мгновенного переключения контекста между трейдингом и психологией.
- **ETL-пайплайн (Часть 2)** с его семантической сегментацией и LLM-классификацией (Таблица 2) обеспечивает *качество* данных и *точность* фильтрации, которые требуются Qdrant.
- **Архитектура Промптов (Часть 3)** является "мозгом", который использует эту инфраструктуру для реализации инновационных механик "Socratic RAG" и "СВТ RAG" , а также проактивного вмешательства.

Следующие Шаги Следующим критическим шагом является создание комплексного оценочного набора данных (evaluation dataset). Этот набор должен быть создан вручную или с помощью LLM и должен содержать пары "вопрос-эталонный ответ". Он должен тестировать не только *релевантность извлечения* (Retrieval effectiveness) и *качество генерации* (Relevance of responses) , но и, что более важно, *корректность переключения персон* и *адекватность срабатывания* проактивных триггеров и этических ограничителей.

Источники

1. A Cheat Sheet and Some Recipes For Building Advanced RAG - Llamaindex, <https://www.llamaindex.ai/blog/a-cheat-sheet-and-some-recipes-for-building-advanced-rag-803a9d94c41b>
2. Qdrant Documentation, <https://qdrant.tech/documentation/>
3. What is a Vector Database? - Qdrant, <https://qdrant.tech/articles/what-is-a-vector-database/>
4. Built for Vector Search - Qdrant, <https://qdrant.tech/articles/dedicated-vector-search/>
5. Combining Vector Search and Filtering - Qdrant, <https://qdrant.tech/course/essentials/day-2/filterable-hnsw/>
6. Payload - Qdrant, <https://qdrant.tech/documentation/concepts/payload/>
7. The Comprehensive Guide to Vector Databases and Qdrant: From Theory to Production, <https://scaibu.medium.com/the-comprehensive-guide-to-vector-databases-and-qdrant-from-theory-to-production-ced44e4ae579>
8. Qdrant - Docs by LangChain, <https://docs.langchain.com/oss/python/integrations/vectorstores/qdrant>
9. Filtering - Qdrant, <https://qdrant.tech/documentation/concepts/filtering/>
10. Indexing - Qdrant, <https://qdrant.tech/documentation/concepts/indexing/>
11. A Complete Guide to Filtering in Vector Search - Qdrant, <https://qdrant.tech/articles/vector-search-filtering/>
12. Project: Building a Semantic Search Engine - Qdrant, <https://qdrant.tech/course/essentials/day-1/pitstop-project/>
13. Taxonomy - Technical Analysis, <http://www.technicalanalysis.org.uk/taxonomy.html>
14. 40

Topics List | PDF - Scribd, <https://www.scribd.com/document/825423607/40-Topics-List> 15. Best Practices in RAG Evaluation: A Comprehensive Guide - Qdrant, <https://qdrant.tech/blog/rag-evaluation-guide/> 16. Setup Hybrid Search with FastEmbed - Qdrant, <https://qdrant.tech/documentation/beginner-tutorials/hybrid-search-fastembed/> 17. Hybrid Search Revamped - Building with Qdrant's Query API, <https://qdrant.tech/articles/hybrid-search/> 18. youtube-transcript-api - PyPI, <https://pypi.org/project/youtube-transcript-api/> 19. How to extract Youtube Video transcripts using Youtube API on Python - Stack Overflow, <https://stackoverflow.com/questions/76856230/how-to-extract-youtube-video-transcripts-using-youtube-api-on-python> 20. Analyzing Youtube Transcripts with Instructor, <https://python.useinstructor.com/blog/2024/07/11/youtube-transcripts/> 21. What's the difference between youtube-dl and yt-dlp?? : r/youtubedl - Reddit, https://www.reddit.com/r/youtubedl/comments/1cu2r1u/whats_the_difference_between_youtube_dl_and_ytdlp/ 22. Youtube-dl vs yt-dlp : r/youtubedl - Reddit, https://www.reddit.com/r/youtubedl/comments/mlxxxm/youtubedl_vs_ytdlp/ 23. Using yt-dlp to download youtube transcript | by jon allen - Medium, <https://medium.com/@jallenswrx2016/using-yt-dlp-to-download-youtube-transcript-3479fccad9ea> 24. Transcribing any Youtube Video with Python: A Step-by-Step guide | by Naveen Malla, <https://naveen-malla.medium.com/transcribing-any-youtube-video-with-python-a-step-by-step-guide-cea7dd4e32f5> 25. Chunking Strategies for LLM Applications - Pinecone, <https://www.pinecone.io/learn/chunking-strategies/> 26. Text splitters - Docs by LangChain, <https://docs.langchain.com/oss/python/integrations/splitters> 27. Trading Psychology: A Therapeutic Guide to Master Your Mind, <https://centrepsicologijaumeprimer.com/en/trading-psychology-guide/> 28. Identifying Chart Patterns with Technical Analysis - Fidelity Investments, https://www.fidelity.com/bin-public/060_www_fidelity_com/documents/learning-center/Identifying-Chart-Patterns.pdf 29. Best Chunking Strategies for RAG in 2025 - Firecrawl, <https://www.firecrawl.dev/blog/best-chunking-strategies-rag-2025> 30. langchain_experimental.text_splitter.SemanticChunker — LangChain 0.2.17, https://api.python.langchain.com/en/latest/text_splitter/langchain_experimental.text_splitter.SemanticChunker.html 31. Exploring the Frontiers of Text Segmentation for Better RAG Systems - Medium, <https://medium.com/@hajar.mousannif/exploring-the-frontiers-of-text-segmentation-for-better-rag-systems-1b9f9af49258> 32. How to Write Effective System Prompts for AI Assistants: Complete Guide, <https://systempromptmaster.com/learning/system-prompt-writing-guide> 33. Guide to Writing System Prompts: The Hidden Force Behind Every AI Interaction - Sahara AI, <https://saharaai.com/blog/writing-ai-system-prompts> 34. Prompt Engineering Patterns for Successful RAG Implementations - Shittu Olumide Ayodeji, <https://iamholumeedey007.medium.com/prompt-engineering-patterns-for-successful-rag-implementations-b2707103ab56> 35. Exploring emotional patterns in social media through NLP models to unravel mental health insights - PMC - NIH, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11730989/> 36. Machine Learning for Identifying Emotional Expression in Text: Improving the Accuracy of Established Methods - PMC - PubMed Central, <https://pmc.ncbi.nlm.nih.gov/articles/PMC7467127/> 37. Add Depth to AI Responses with Emotional Language - Learn Prompting, https://learnprompting.org/docs/advanced/zero_shot/emotion_prompting 38. RAG vs fine-tuning vs. prompt engineering - IBM, <https://www.ibm.com/think/topics/rag-vs-fine-tuning-vs-prompt-engineering> 39. System Prompt

Engineering in Gen AI Applications - Belitsoft,
<https://belitsoft.com/system-prompt-engineering-in-gen-ai-applications> 40. Advanced Prompt Engineering Techniques: Examples & Best Practices - Patronus AI,
<https://www.patronus.ai/lm-testing/advanced-prompt-engineering-techniques> 41. Effective Prompts for AI: The Essentials - MIT Sloan Teaching & Learning Technologies,
<https://mitsloanedtech.mit.edu/ai/basics/effective-prompts/> 42. Use of generative AI chatbots and wellness applications for mental health,
<https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-chatbots-wellness-apps> 43. Building a Mental Health Chatbot in 2025 | The Ultimate Guide - Topflight Apps,
<https://topflightapps.com/ideas/build-mental-health-chatbot/> 44. Designing Chatbots to Treat Depression in Youth: Qualitative Study - PMC - PubMed Central,
<https://pmc.ncbi.nlm.nih.gov/articles/PMC12199846/> 45. 35+ Powerful CBT Exercises & Techniques for Therapists - Positive Psychology,
<https://positivepsychology.com/cbt-cognitive-behavioral-therapy-techniques-worksheets/> 46. Ep. 24: Dr. Brett Steenbarger — Using the Tools of CBT to Be a Better Trader - Seth Gillihan,
<https://sethgillihan.com/ep-24-dr-brett-steenbarger-using-the-tools-of-cbt-to-be-a-better-trader/> 47. Financial Stress and Anxiety Resources - Crisis Text Line,
<https://www.crisistextline.org/topics/financial-stress-and-anxiety-resources/> 48. Coping with Financial Stress - HelpGuide.org,
<https://www.helpguide.org/mental-health/stress/coping-with-financial-stress> 49. Overcoming FOMO & Revenge Trading in Forex – Why Patience Pays - ACY Securities,
<https://acy.com/en/market-news/education/market-education-overcoming-fomo-revenge-trading-j-o-20250828-100501/> 50. Can CBT Principles Make You a Better Trader? - Psychology Today,
<https://www.psychologytoday.com/us/blog/think-act-be/201901/can-cbt-principles-make-you-a-better-trader> 51. Trading Psychology Dealing With FOMO and Revenge Trading - FundYourFX,
<https://fundyourfx.com/trading-psychology-dealing-with-fomo/> 52. Using generic AI chatbots for mental health support: A dangerous trend - APA Services,
<https://www.apaservices.org/practice/business/technology/artificial-intelligence-chatbots-therapists> 53. How to Deal with Financial Stress - The Jed Foundation,
<https://jedfoundation.org/resource/how-to-deal-with-financial-stress/> 54. Coping with financial worries - NHS,
<https://www.nhs.uk/mental-health/advice-for-life-situations-and-events/how-to-cope-with-financial-worries/> 55. Detection of emotion by text analysis using machine learning - Frontiers,
<https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1190326/full> 56. Mastering FOMO and Revenge Trading: A Trader's Guide,
<https://mavericktrading.com/mastering-fomo-and-revenge-trading-a-traders-guide/> 57. Complete Guide to Investing & Trading Psychology (Updated 2025),
<https://synapsetrading.com/trading-guides/trading-psychology/> 58. Questioning the Bot: Socratic Chats with AI - The AI Pedagogy Project,
<https://aipedagogy.org/assignment/questioning-the-bot-socratic-chats-with-ai-2/> 59. Socratic Questioning with Chat GPT - Secondary School Lesson Demonstration - YouTube,
<https://www.youtube.com/watch?v=fAeYkunflq4> 60. Enhancing Critical Thinking in Education by means of a Socratic Chatbot - arXiv, <https://arxiv.org/html/2409.05511v1> 61. What Is Risk Management in Finance, and Why Is It Important? - Investopedia,
<https://www.investopedia.com/terms/r/riskmanagement.asp> 62. The Art of Writing Great System Prompts | by Saurabh Singh | Towards Dev - Medium,
<https://medium.com/towardsdev/the-art-of-writing-great-system-prompts-abb22f8b8f37> 63. How to Write AI Prompts for Customer Service (With Examples) - Talkative,

<https://gettalkative.com/info/ai-prompts-for-customer-service> 64. Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions - arXiv, <https://arxiv.org/html/2403.01222v1>
65. Mastering Trading Psychology - Bear Bull Traders,
<https://bearbulltraders.com/wp-content/uploads/pdfs/books/Mastering-Trading-Psychology.pdf>