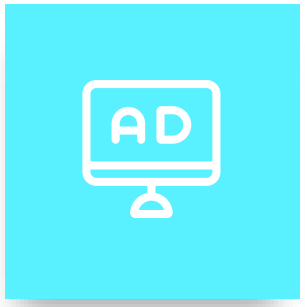
Использование системы из нескольких ИИ – агентов для снижения риска галлюцинаций

Олег Зельдин Апекс Берг

Использование системы из нескольких ИИ – агентов для снижения риска галлюцинаций

Обзор



Аннотация

Предлагается мультиагентная система, обрабатывающая клиентские запросы по SMS. Она сочетает LLM-агентов с нечеткой логикой, чтобы повысить качество сервиса и скорость ответа, одновременно уменьшая риск «галлюцинаций» моделей. Цель — улучшить лояльность клиентов и поддержать рост доли рынка.

Источник

Название статьи и дата публикации

- Abd Elrahman Amer and Magdi Amer. Using multi-agent architecture to mitigate the risk of LLM hallucinations. 02.07.2025

Ссылка

- https://clck.ru/3PxVQo

Риски финансовых потерь при неудачном использовании GenAI

Ошибки бота могут приводить к компенсациям недовольным клиентам, упущенным продажам или даже штрафам регуляторов.

«Пока громких скандалов, повлекших штрафы, не было, но компании заранее закладывают риск-резервы на возможные инциденты, страхуют ответственность. «

Об этом я говорил всего полгода назад... Однако...

Галлюцинации AI и их последствия

Галлюцинации AI – это уверенное изложение неправды. Это приводит к реальным рискам для клиентов и бизнеса. Ниже — показательные случаи, из-за которых компании осторожничают с «LLM-only» сервисами и переходят к более надёжным схемам (мультиагенты, правила, нечёткая логика). В мультиагентных системах риск галлюцинаций возрастает (каскадный эффект)*

✈️ Кейс авиакомпании (Канада): суд признал компанию ответственной за ложный совет чат-бота клиенту.

🛡️ Кейс страховой (США): бот дал неверную рекомендацию; клиенту сначала отказали, но после огласки компания выплатила сумму, «как сказал бот».

⚖️ Кейсы в судах (США/Канада): юристы подали документы, сгенерированные LLM, где цитировались несуществующие судебные прецеденты → предупреждения и санкции.

📉 Вывод: без контролируемой логики и проверок фактов LLM-сервисы несут правовые и репутационные риски; потому в работе предлагается мультиагентная архитектура с нечёткой логикой для снижения «галлюцинаций».

* Guo, T., Chen, X., Wang, Y., Chang , R., Pei , S., Chawla, N. V., . . . Zhang, X. (2024). Large Language Model based Multi-Agents: A Survey of Progress and Challenges. *Proceedings of the Thirty-Third International Joint Conference onArtificial Intelligence, {IJCAI-24}* (pp. 8048-8057). Jeju, South Korea: International Joint Conferences on Artificial Intelligence Organization.
doi: https://doi.org/10.24963/ijcai.2024/890

Общее описание кейса

Поликлиника рассылает клиентам автоматические сообщения разного типа предполагающие ответ, например:

- Renewal. Продлить/остановить продление истекающего рецепта (тем, кто регулярно получает лекарства по рецептам)
- Visit-Remind. Напомнить о записи на прием к врачу
- Update. Оповестить о ситуациях, когда возможность посетить врача появляется раньше (если клиент просил об этом)
- Inform. Проинформировать о возможности сдать регулярные анализы в соответствии с медицинской картой

Проблематика (на примере обновления рецептов)

Клиентам предлагается в ответ отправить код действия (R – продлить, S – остановить продление) и код лекарства. Разным лекарствам назначаются разные коды. Клиенты могут комбинировать несколько кодов в одном сообщении, указывая, какие препараты они хотят продлить, а какие — остановить.

На практике, коды могут указываться разными способами, в разном порядке и т.п. Также, кроме кодов, клиенты часто добавляют предложения с особыми инструкциями или вопросами, типа:

«Ещё перестаньте продлевать витамины»

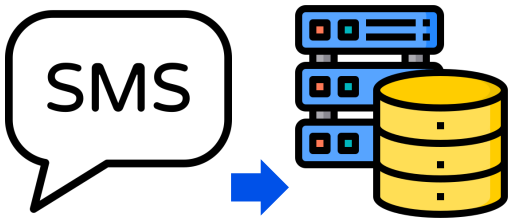
«И у сиропа вкус странный»

«Можно записаться на прививку в воскресенье?».

Традиционные технологии не справляются с обработкой таких сообщений, поскольку приложению трудно корректно интерпретировать намерения клиента. Использование способности LLM понимать естественный язык подходит для повышения успешности обработки сообщений.

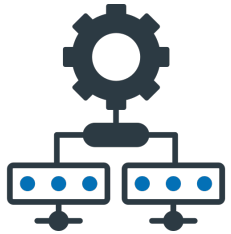
1-й этап. Обработка контента сообщений, выбор фронт-агента и подготовка информации для Shared Message Pool

Сервис обработки входящих СМС



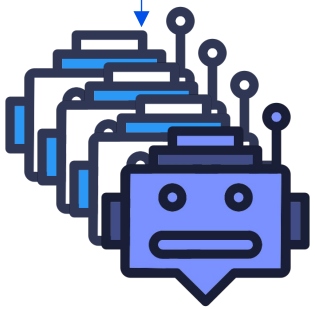
1. Аутентификация клиента
2. Запись в БД метаданных, в т.ч. ТИП сообщения:
   - RENEWAL
   - VISIT-REMIND
   - UPDATE
   - INFORM

Агент-диспетчер



1. Читает конфигурацию
2. Проверяет условия
3. Выбирает фронт-агента нужного типа

Фронт-агент



- Renewal Agent (RA)
- Visit-Remind Agent (VRA)
- Update Agent (UA)
- Inform Agent (IA)

1. Читает SMS типа RENEWAL
2. Чистит “вежливости”, предлоги
3. Режет на слова и применяет RegEx для ключевых слов «продлить/остановить».
4. Считает степень уверенности распознавания (для Fuzzy Logic).
5. Формирует результат и публикует в Shared Message Pool со step-id = 1
6. Сохраняет оригинал смс в БД.

Пара слов про Regex (регулярные выражения) и Fuzzy Logic (нечеткая логика)

RegEx

Идея - 1950-е годы (Стивен Клини), популяризация в языках программирования – 1980-90-е)

Зачем:
Чтобы **компактно описывать шаблоны в тексте** и быстро искать/выделять/заменять строки по правилам (электронные логи, парсинг, валидации форм и т. п.). Теоретическая база (регулярные языки ↔ конечные автоматы) дала быстрые и предсказуемые алгоритмы.

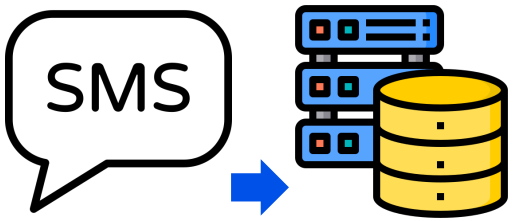
Fuzzy Logic.

Идея – 1965 (Лотфи А. Заде), первое промышленное применение 1970-80-е (Япония – поезда, бытовая техника). 1990-2000-е – массовое применение – (стиральные машины, камеры, климат-контроль, экспертные системы)

Зачем:
Чтобы **моделировать расплывчатые, человеческие понятия** («тёпло», «важно», «слегка») и принимать **плавные решения** там, где бинарной логики мало. Подходит для управления, диагностики, ранжирования, где нужны «степени» вместо жёстких порогов.

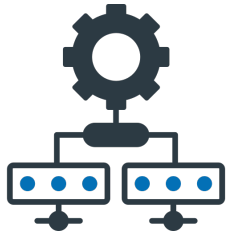
1-й этап. Обработка контента сообщений, выбор фронт-агента и подготовка информации для Shared Message Pool

Сервис обработки входящих СМС



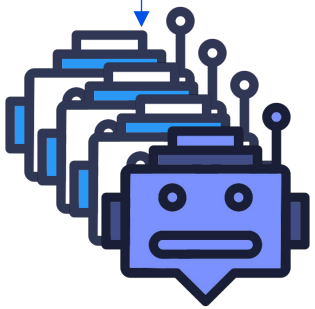
1. Аутентификация клиента
2. Запись в БД метаданных, в т.ч. ТИП сообщения:
   - RENEWAL
   - VISIT-REMIND
   - UPDATE
   - INFORM

Агент-диспетчер



1. Читает конфигурацию
2. Проверяет условия
3. Выбирает фронт-агента нужного типа

Фронт-агент



- Renewal Agent (RA)
- Visit-Remind Agent (VRA)
- Update Agent (UA)
- Inform Agent (IA)

1. Читает SMS типа RENEWAL
2. Чистит “вежливости”, предлоги
3. Режет на слова и применяет RegEx для ключевых слов «продлить/остановить».
4. Считает **степень уверенности распознавания** (для Fuzzy Logic).
5. Формирует результат и **публикует в Shared Message Pool со step-id = 1**
6. Сохраняет оригинал смс в БД.

2-й этап обработки. Shared Message Pool

Что такое?

Одна из 4-х моделей коммуникаций между агентами. «Общий стол» сообщений для всех агентов. Размещают все, но читают – только подписанные на определенные критерии

Зачем?

Развязать агентов: каждый работает независимо по своей скорости и не знает деталей других.

Масштабировать: можно добавлять/дублировать агентов без переделки системы.

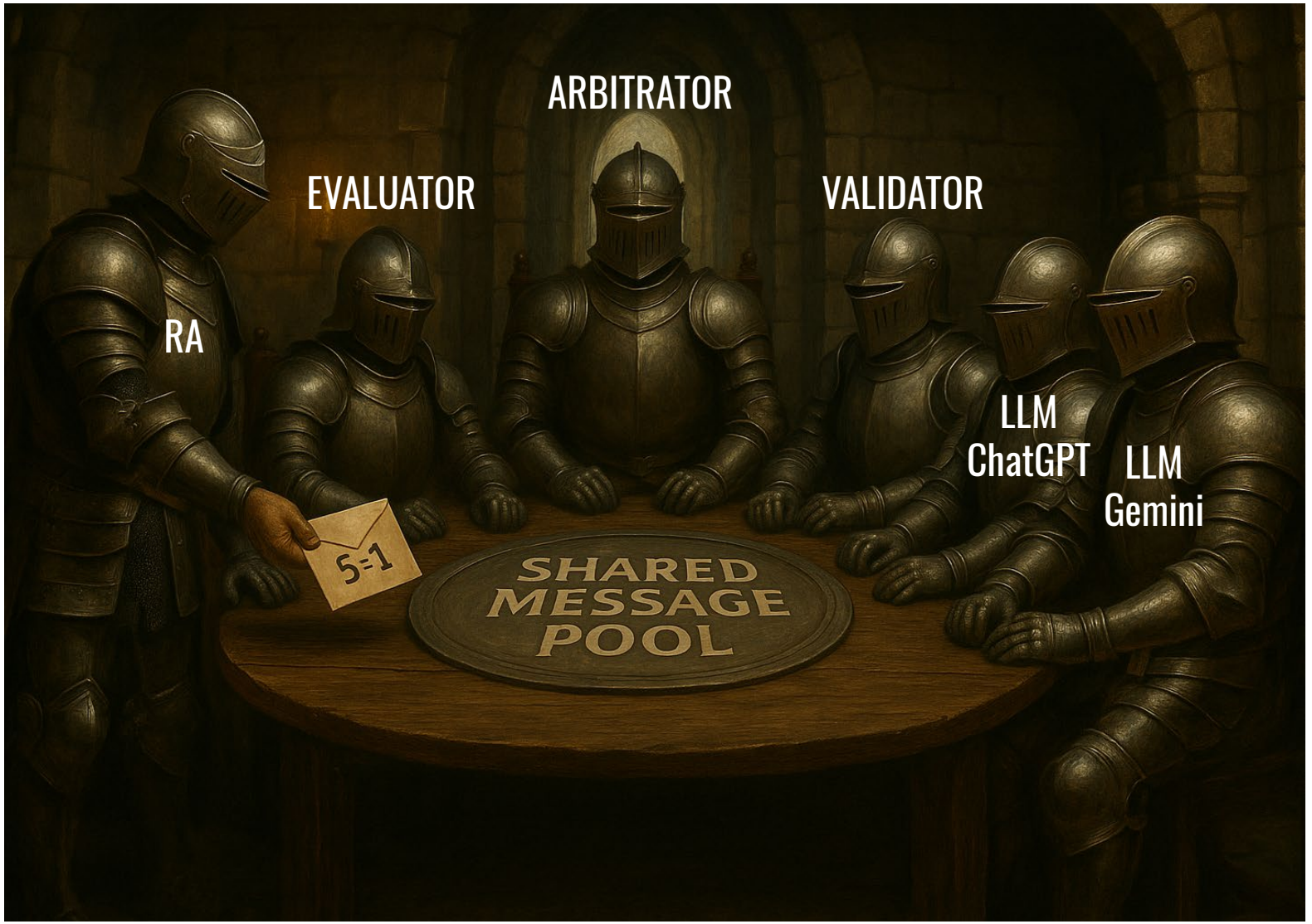
Надёжность: сообщения хранятся, можно повторно доставить при сбоях.

Гибкая маршрутизация: любое правило. Агенты включаются в процесс, когда нужно.

2-й этап обработки. Shared Message Pool

1 → 2 → 3 → 4 → 5 → 6 → 7

1. RA – принес и положил на стол
2. Arbitrator. Самый главный. Определяет правила работы и следит за фиксацией шагов
3. Evaluator – первым читает сообщение, определяет метки
4. LLM – агенты (ChatGPT и Gemini) – работают после Evaluator-а, если так решит Arbitrator
5. VALIDATOR –проверяет LLM и оценивает риск отмены важных лекарств

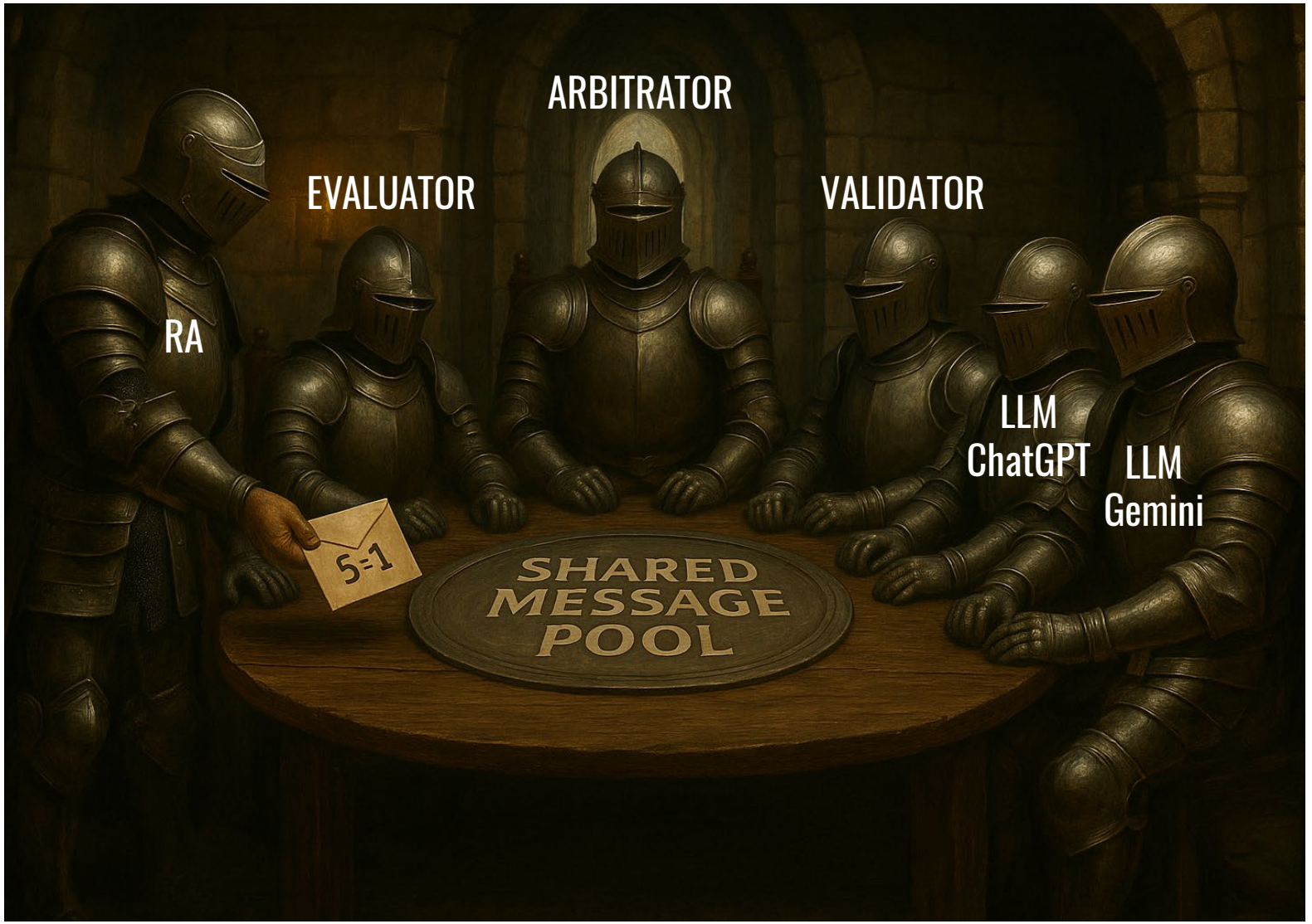


2-й этап обработки. Shared Message Pool

1 → 2 → 3 → 4 → 5 → 6 → 7

EVALUATOR

1. Читает *степень уверенности распознавания* от RA
2. Считает *степень важности клиента* из CRM
3. Думает по установленным правилам (Fuzzy Logic)
4. Отдает на «общий стол» одну из меток:
   - **Выполнить** (продление/отмена)
   - **Отдать в LLM**
   - **Отказ** (продление/отмена)

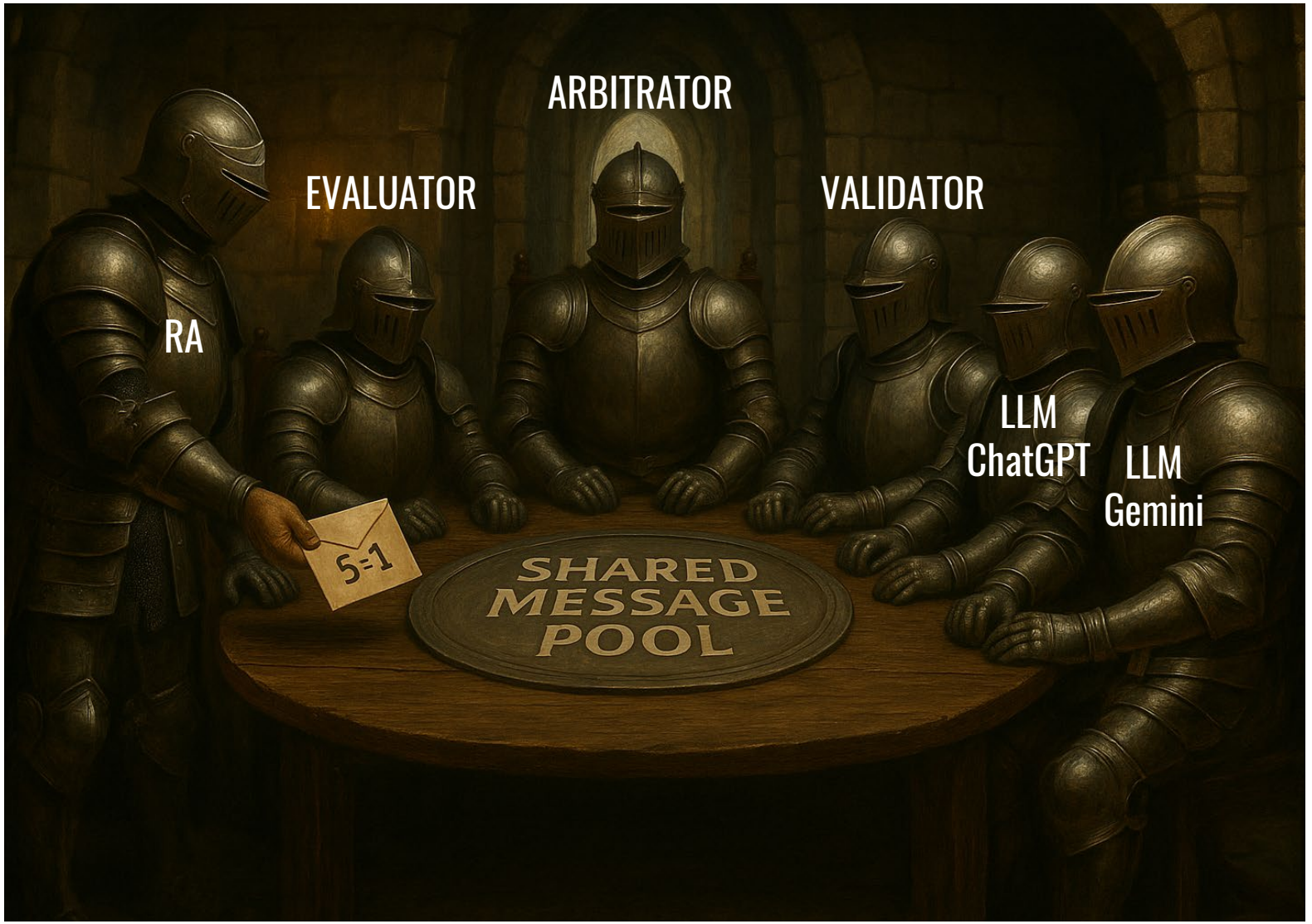


2-й этап обработки. Shared Message Pool

1 → 2 → 3 → 4 → 5 → 6 → 7

ARBITRATOR

1. Читает метку от EVALUATORa
2. Если метка «Выполнить» или «Отказ» – на последний шаг
3. Если метка «Отдать в LLM», присваивает step-id=2 (сигнал к прочтению для LLM)
4. Контролирует учет записей каждого хода в привязке к id события (смс)

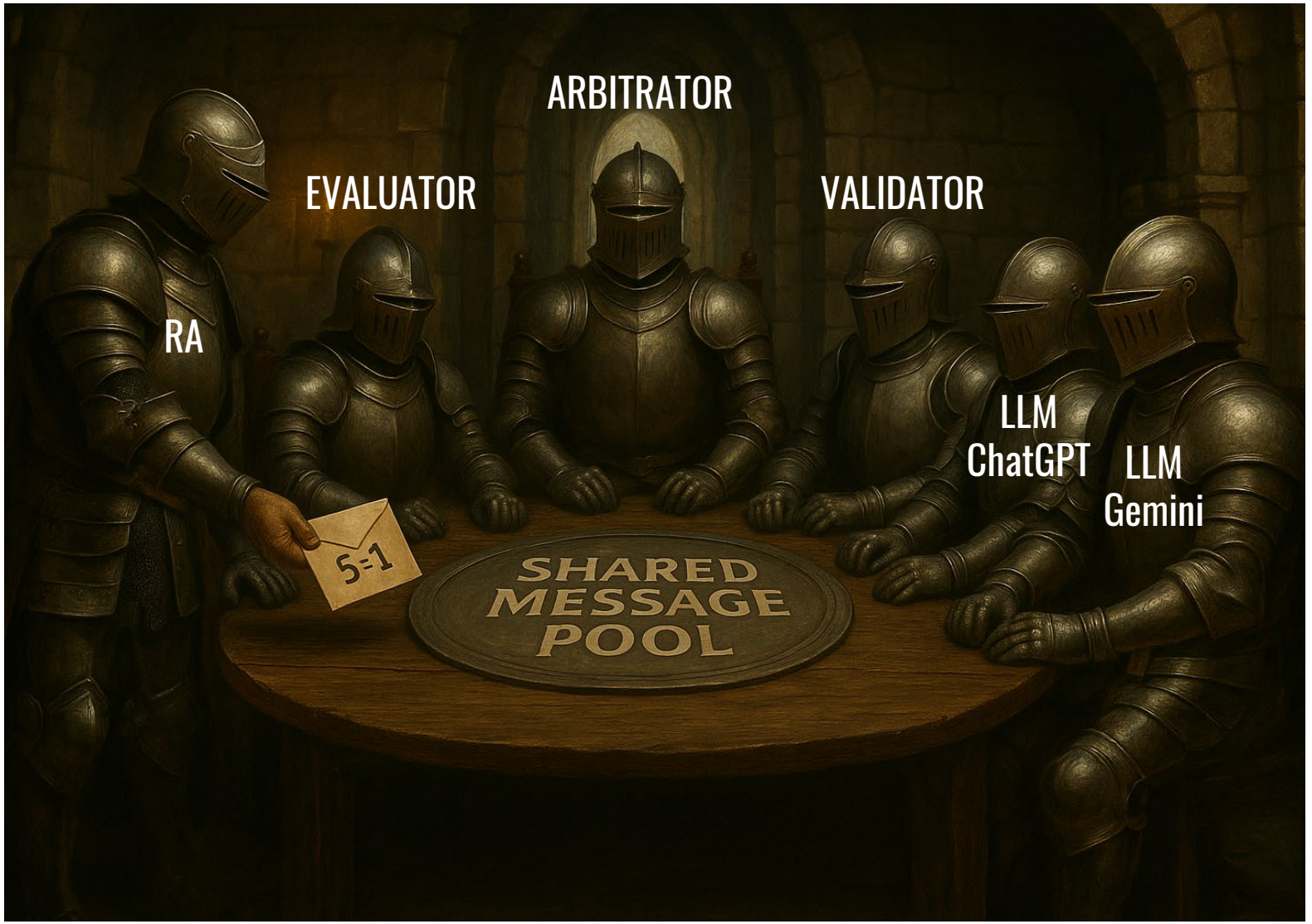


2-й этап обработки. Shared Message Pool

1 → 2 → 3 → **4** → 5 → 6 → 7

LLM ChatGPT и Gemini

1. Независимо читают оригинал сообщения
2. Классифицируют список ключевых слов (независимо от RA и друг от друга)
3. Извлекают список жалоб и вопросов, если найдут

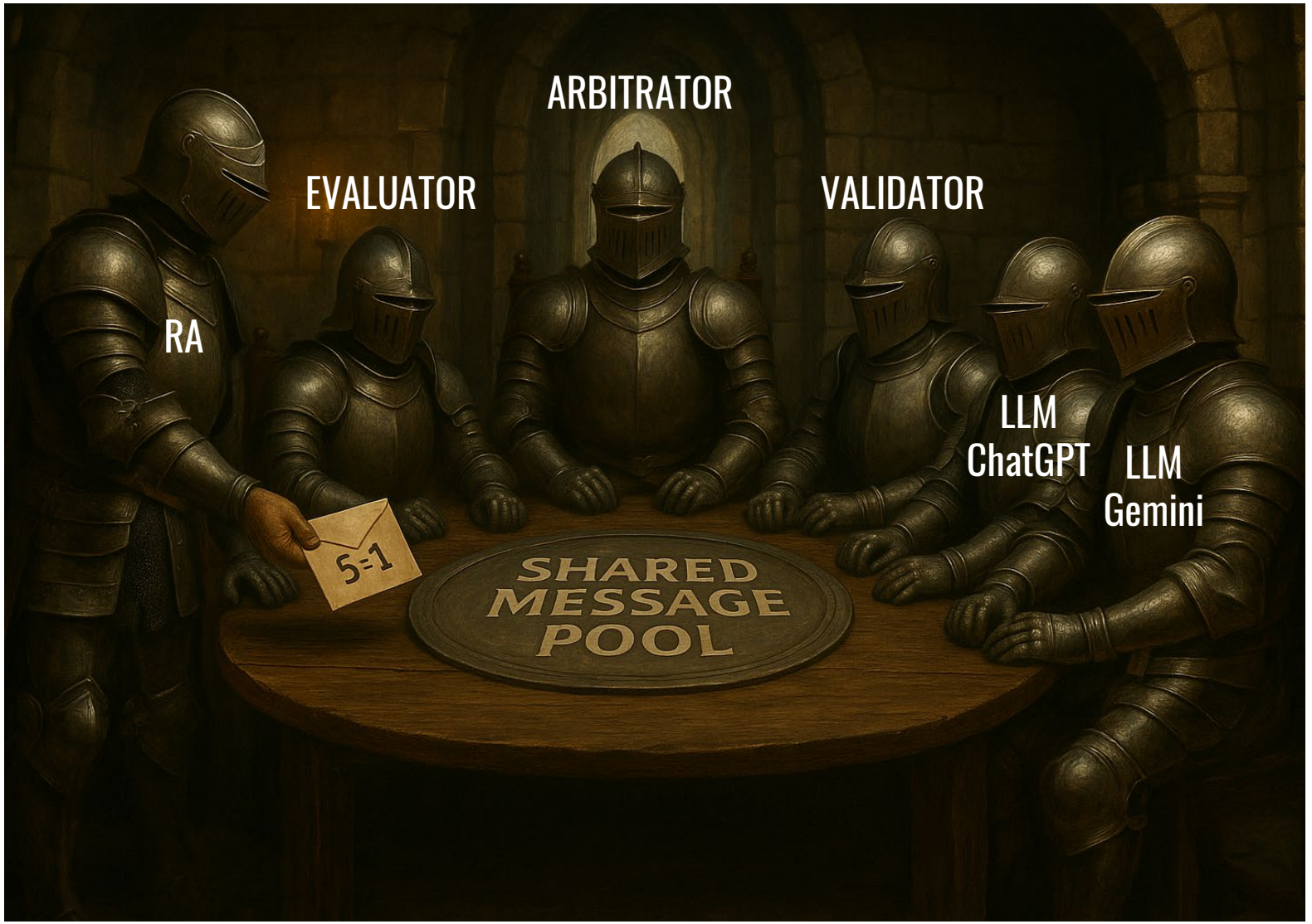


2-й этап обработки. Shared Message Pool

1 → 2 → 3 → 4 → **5** → 6 → 7

ARBITRATOR

1. Направляет результаты работы LLM агентов на VALIDATOR
   - Ключевые слова
   - Жалобы, просьбы
2. Контролирует учет записей каждого хода в привязке к id события (смс)

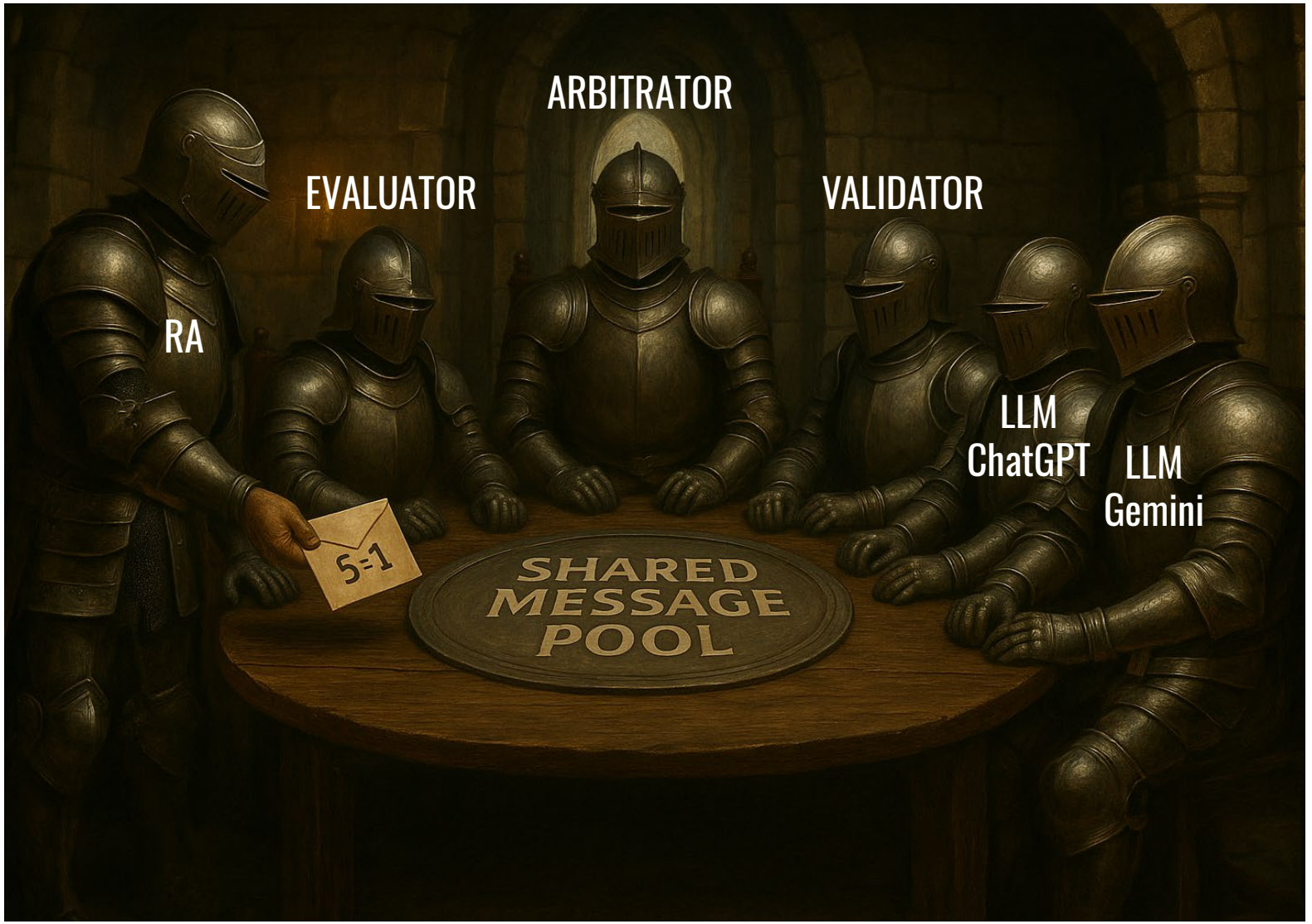


2-й этап обработки. Shared Message Pool

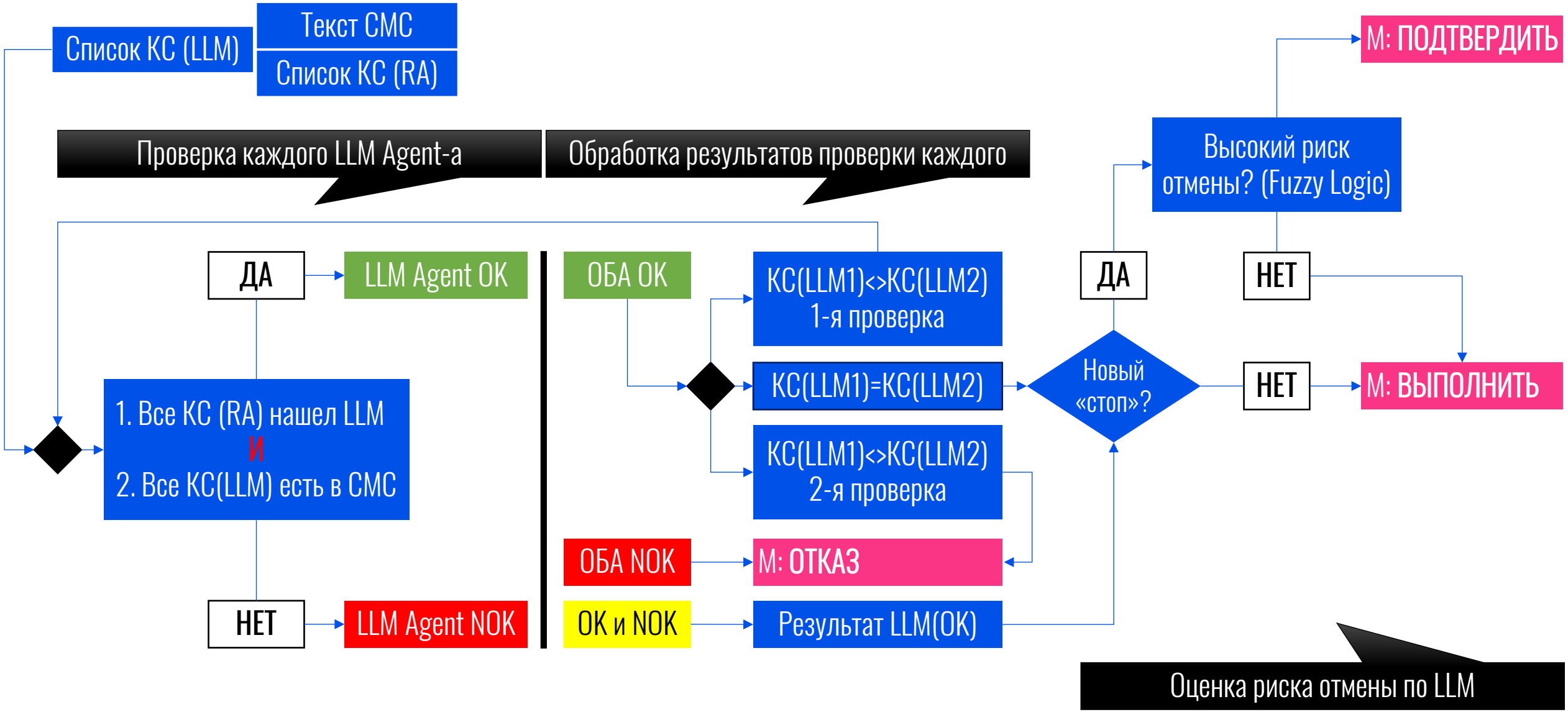
1 → 2 → 3 → 4 → 5 → **6** → 7

VALIDATOR

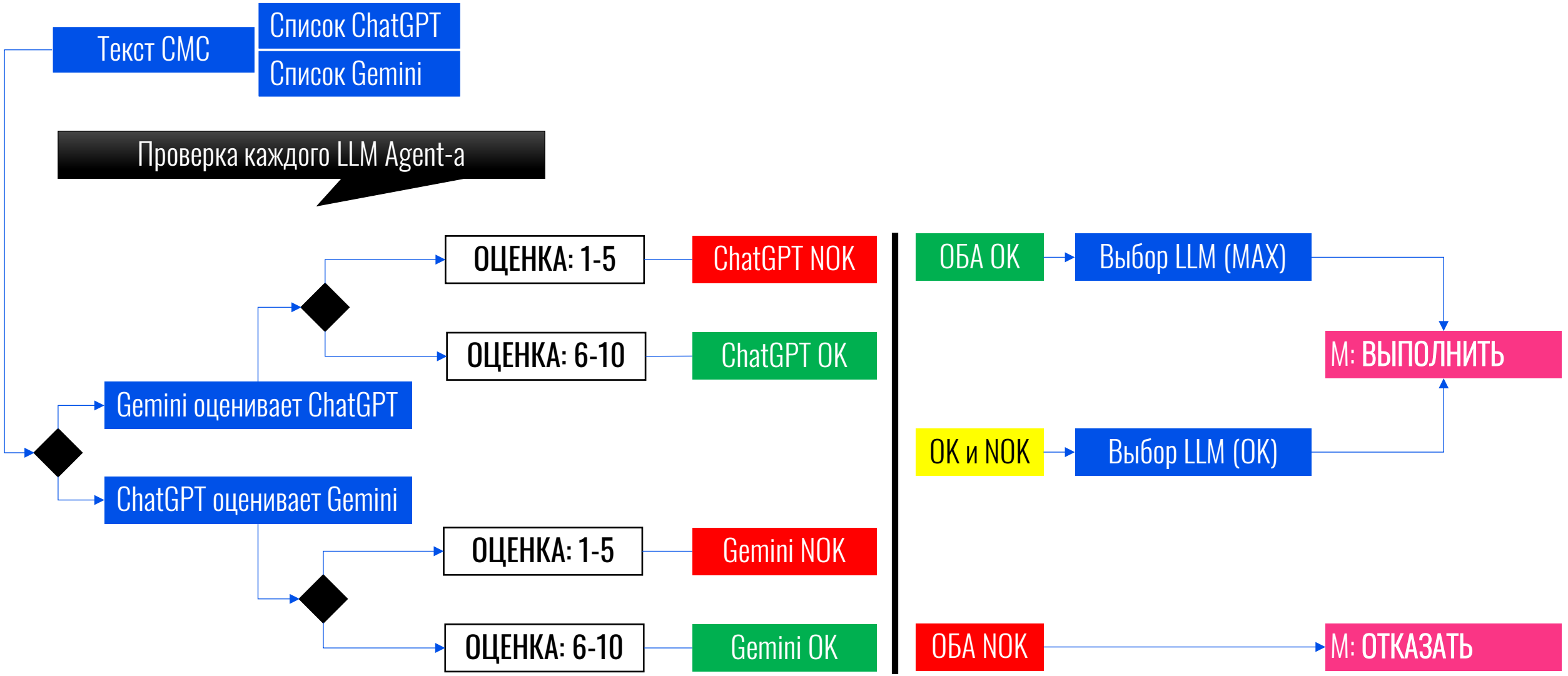
1. Делает проверку на галлюцинации результатов LLM по ключевым словам
2. Если LLM нашли новую команду клиента по остановке важного препарата, оценивает риск остановки приема важного препарата (Fuzzy Logic)
3. Делает проверку на галлюцинации по выявленным жалобам или просьбам
4. Ставит свои метки в событии для действий ARBITRATOR



Проверки VALIDATOR–ом LLM-агентов. Ключевые слова (КС) – продление/отмена



Проверки VALIDATOR–а LLM-агентов. Жалобы/просьбы

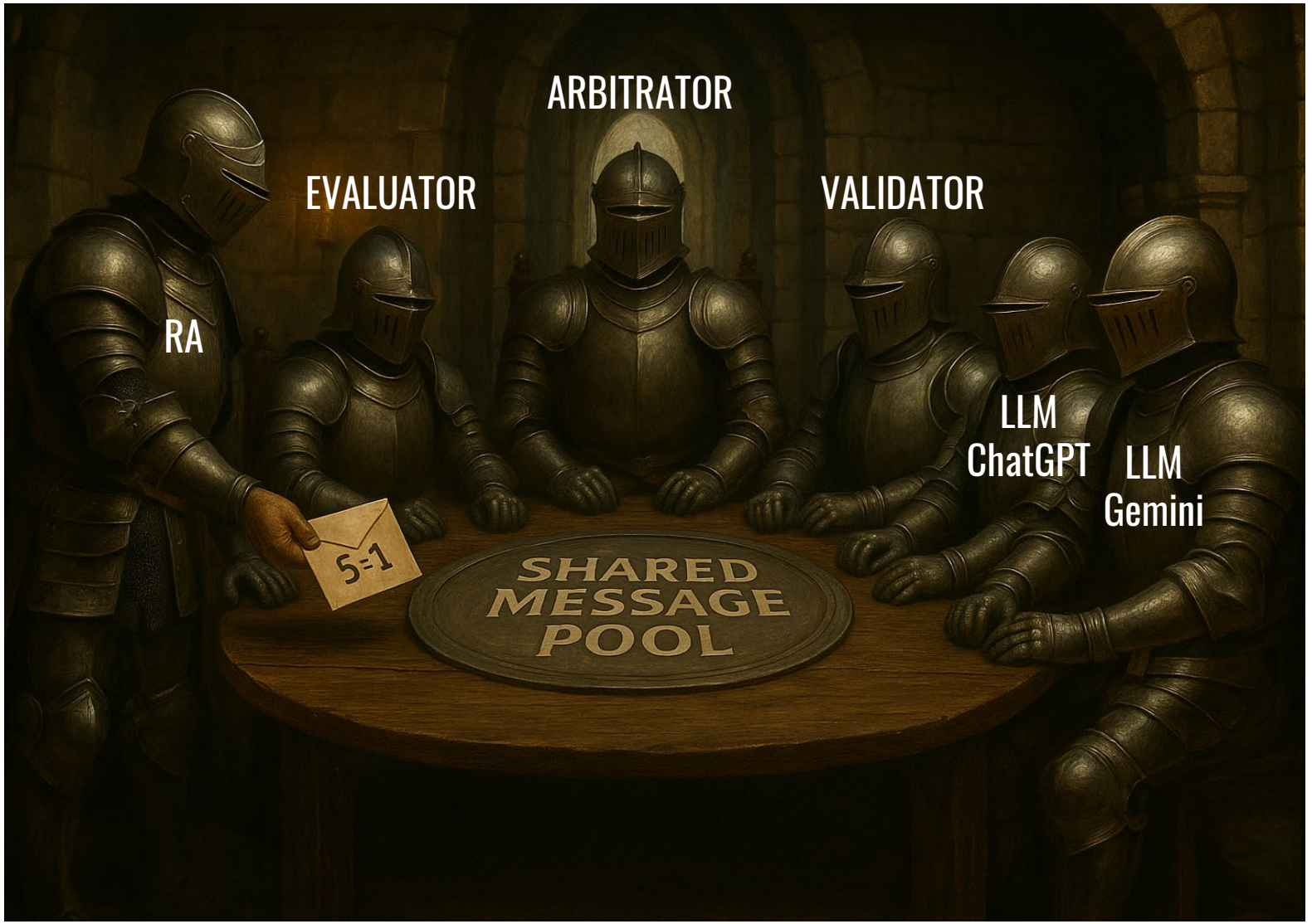


2-й этап обработки. Shared Message Pool. Финал

1 → 2 → 3 → 4 → 5 → 6 → **7**

ARBITRATOR

1. Осуществляет действие согласно меток от EVALUATOR и VALIDATOR
2. Контролирует учет записей каждого хода в привязке к id события (смс)



Действия по меткам

M: ВЫПОЛНИТЬ (Продление/отмена) → End Point: Продлить/отменить

M: ВЫПОЛНИТЬ (Жалоба/просьба) → End Point: Маршрутизация эксперту

M: ОТКАЗ (Жалоба/просьба) → SMS клиенту: Обратитесь в поддержку

M: ОТКАЗ (Продление/отмена) → SMS клиенту: Обратитесь в поддержку

M: ПОДТВЕРДИТЬ (отмена) → SMS клиенту: Подтвердите отмену

Результат

1. Это POC (Proof of Concept)
2. Предварительная проверка на 60 сообщениях показала удовлетворительные результаты
3. Авторы планируют усовершенствовать схему, провести более обширное тестирование и отчитаться

Система MindFlow – первый открытый мультимодальный агент на базе LLM, нацеленный на сферу поддержки клиентов в e-commerce

Обзор

Источник

Название статьи и дата публикации

- MindFlow: Revolutionizing E-commerce Customer Support with Multimodal LLM Agents. 24.03.2025

Ссылка

- https://clck.ru/3Q2k2A

Первый бенчмарк ECom-Bench для оценки мультимодальных LLM-агентов в задачах поддержки клиентов интернет-магазинов

Обзор

Источник

Название статьи и дата публикации

- ECom-Bench: Can LLM Agent Resolve Real-World E-commerce Customer Support Issues?. 24.03.2025

Ссылка

- https://clck.ru/3Q2jy5

СПАСИБО ЗА ВНИМАНИЕ!

apexberg.ru

ТГ-КАНАЛ:
Клиентский сервис –
искусство служить людям