Агентно-ориентированная модель обработки логов, основанная на синтезе методов промпт-инжиниринга и цепочки рассуждений

Пашигорев К.И., приглашенный преподаватель,

Базовая кафедра Сбербанка

Гипотезы

Гипотеза 1: Генерацию новости и сокращение количества атрибутов логов при сохранении качества новости возможно осуществить итеративным применением LLM

Гипотеза 2: при итеративном применении LLM режим работы с рассуждениями позволяет повысить эффективность этапа сокращения логов при сохранении качества новости

Ограничения

1. Отсутствует доступ к изменению архитектуры LLM
2. Отсутствует возможность дообучения модели
3. Исключить применение RAG

Разработка и применение мультиагентных систем в корпоративной среде

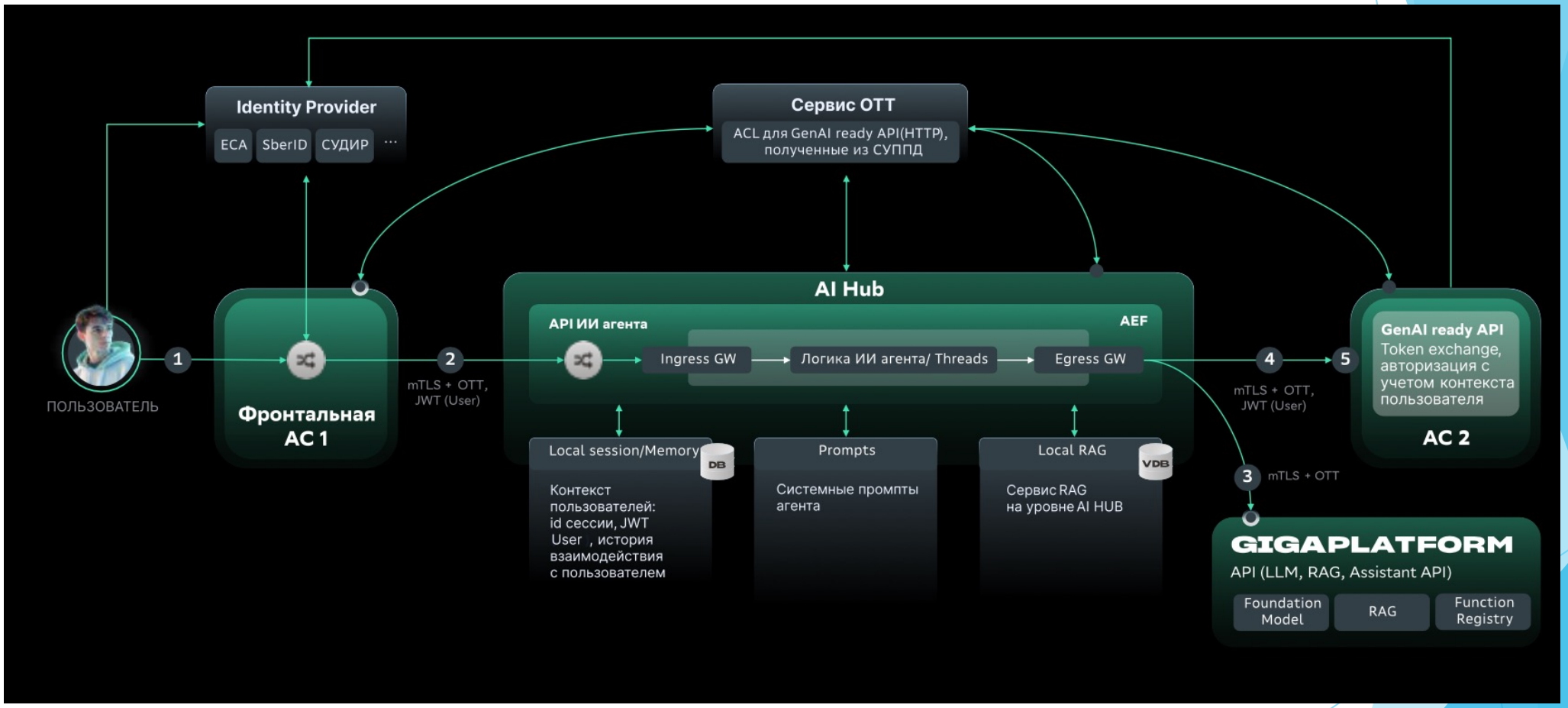

Рисунок 1. Изолированная среда исполнения кода

Источник: гайд "Разработка и применение мультиагентных систем в корпоративной среде https://b2b.giga.chat/multi-agent-system

Пример датасета

	Название дата-продукта	Дата-продукт	Категория в СМД	Тип дата-продукта	Кластер	Схема БД	Менеджеры данных ДП	uuid
0	Витрина Суммаризация д	СВД ДРПА. DS. Витрина Суммаризаі	Пользовательские данные	Специализированная витрина	A	CUSTOM_DRPA_DS_DIALOGUE_SUM	Гуса	0A0A2
1	-	-	-	-	-	-	-	0A0CD
2	Очная аутентификация в	БХБ160 ММБ Очная аутентификаци	Реплики ППРБ	Реплика	S	PLATFORM_F2FAUTH	Вак	0A0CD
3	Очная аутентификация в	БХБ160 ММБ Очная аутентификаци	Реплики ППРБ	Реплика	S	PLATFORM_F2FAUTH_HIST	Вак	0A0CD
4	БВД HR Структура функці	БВД HR Структура функциональных	Non-Hadoop Пользовательские,	Специализированная витрина	C	S_GRNPLM_VD_HR_EDP_VD	Кон	0A1A5
5	[CVRS2.IS1101] Конверси	СВД ФБ [CVRS2.IS1101] Конверсии	Non-Hadoop Пользовательские,	Специализированная витрина	G	S_GRNPLM_AS_FIN_MIS3_MISPLT_IS	Мар	0A1CB
6	СВД БТ Сверточная гексаг	СВД БТ Сверточная гексагональная	Пользовательские данные	Специализированная витрина	A	CUSTOM_T_GDM_CONVOLUTIONAL	-	0A1D1
7	-	-	-	-	-	-	-	0A1D9
8	-	-	-	-	-	-	-	0A1D4
9	СВД БТ Сверточная расш	СВД БТ Сверточная расширенная ге	Пользовательские данные	Специализированная витрина	A	CUSTOM_T_GDM_CONVOLUTIONAL	-	0A1D8
10	SberInfra ПорталДИ Harm	harmonizer	Реплики self-service	Реплика	S	SELFSERVICE_HARMONIZER	Аск	0A1D2
11	СВД БТ Сверточная гексаг	СВД БТ Сверточная гексагональная	Пользовательские данные	Специализированная витрина	A	CUSTOM_T_GDM_CONVOLUTIONAL	-	0A1D3
12	-	-	-	-	-	-	-	0A1DA
13	-	-	-	-	-	-	-	0A1DA
14	СВД БТ Сверточная расш	СВД БТ Сверточная расширенная ге	Пользовательские данные	Специализированная витрина	A	CUSTOM_T_GDM_CONVOLUTIONAL	-	0A1DB
15	-	-	-	-	-	-	-	0A1DC
16	СВД БТ Гексагональная с	СВД БТ Гексагональная сверточная	Пользовательские данные	Специализированная витрина	A	CUSTOM_T_GDM_CONVOLUTIONAL	Куз	0A1DC
17	-	-	-	-	-	-	-	0A1DC
18	СВД БТ Гексагональная с	СВД БТ Гексагональная сверточная	Пользовательские данные	Специализированная витрина	A	CUSTOM_T_GDM_CONVOLUTIONAL	Куз	0A1DE
19	-	-	-	-	-	-	-	0A1DE
20	Ежедневные данные КЮЛ	СВД ФБ [LCC2.IS0958] Кредиты ЮЛ	Non-Hadoop Пользовательские,	Специализированная витрина	G	S_GRNPLM_AS_FIN_MIS3_MISPLT_IS	Мар	0A2D8

Архитектура системы

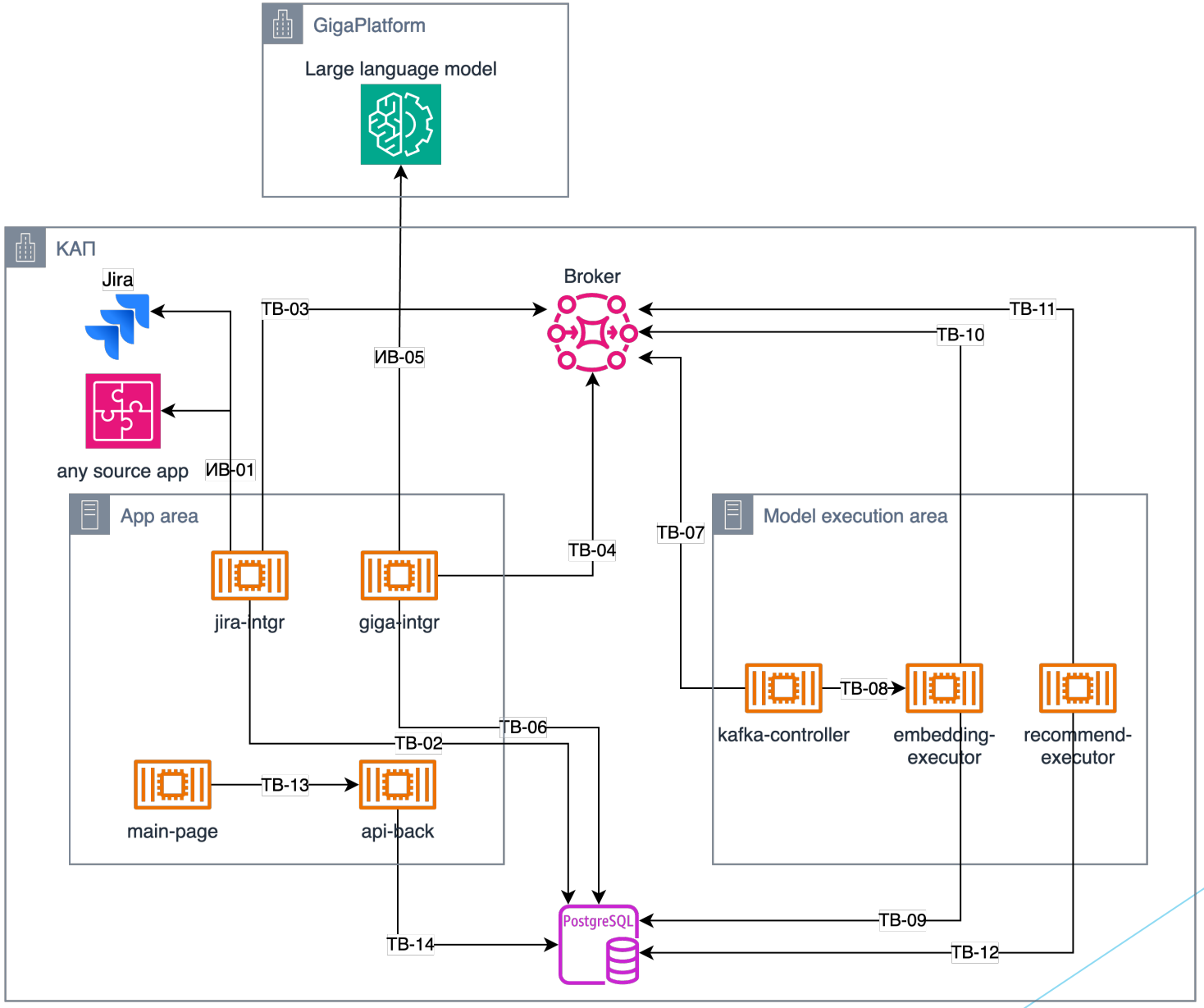

Рисунок 2. Компонентная диаграмма системы

Этапы

Этап 1: Получение ID сырой новости

- Kafka-топик: KAFKA_TOPIC_SOURCE (environment переменная)
- Получение данных по ID из источника (iira_source / any_source)
- Логирование входных данных или ошибок

Этап 2: Трансформация сырых данных

- Преобразование JSON в словарь key-value
- Логирование входных и выходных данных

Этап 3: Фильтрация данных

- Обогащение пользовательского промпта системным промптом для фильтрации
- Запрос в LLM, проверка корректности результата
- Сохранение результата в схему БД data_announcement.cleaned_data
- Логирование данных этапа

Этап 4: Генерация текста новости

- Формирование запроса с системным промптом генерации текста
- Обработка больших текстов (склейка при необходимости)
- Сохранение текста в схему БД data_announcement.announcement_text
- Логирование данных этапа

Этап 5: Фактчекинг

- Добавление пользовательского промпта и системного промпта для проверки фактов. Запрос в LLM, сохранение результата в схему БД data_announcement.fact_check_result
- Логирование входных и выходных данных этапа

Этап 6: Расчет рекомендаций

- Расчет эмбеддингов новостей
- Расчет эмбеддингов пользователей
- Расчет косинусного сходства новости и пользователя
- Логирование входных и выходных данных этапа

Промпты

Промпт	Назначение
Твоя задача определить полезные ключи из переданного списка. Надо избавиться от идентификационных номеров, flags и tags. Полезными считаем ключи, связанные с продуктом, людьми или ключевыми датами. **Подробно по порядку рассуждай, подходит ли ключ в формате: \"Ключ "название" является "продуктом, людьми, датой*, решено "принимаем или отвергаем*, "почему*\".** В самом конце выведи все принятые таким образом ключи в формате \"["ключ1\", …)"	Оптимизация ключей
На основе переданного списка словарей определи ключи, не являющиеся флагами и по которым значения не содержат нечитабельных для человека хэшей, тегов и id. **Максимально подробно описывай этапы проверки каждого ключа в формате: \"рассматриваемый ключ, его значение, подходит ли под критерии или нет, почему\".** Выбери максимум {output size} репрезентативных ключей, не больше! В конца напечатай весь подходящий список в формате \"("key\", …)".	Оптимизация ключей
Сгенерируйте краткую и информативную новостную статью, резюмирующую ключевые детали события, описанного в предоставленных технических данных. Сосредоточьтесь на представлении фактической информации в нейтральном тоне, избегая любых спекуляций, предположений или эмоционального языка. Используйте четкую и логичную структуру и избегайте сенсационного или привлекающего внимание языка. Целью является предоставление объективного резюме события, чтобы читатели могли быстро понять, что произошло.	Генерация текста

Количество ключей

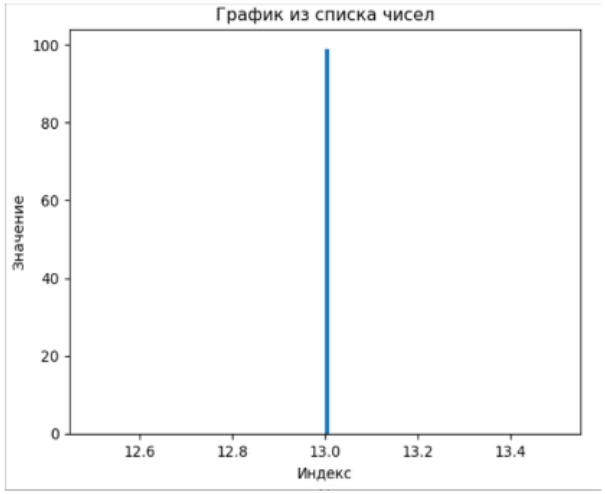

Рисунок 3. Количество ключей для промпта №1 без рассуждения

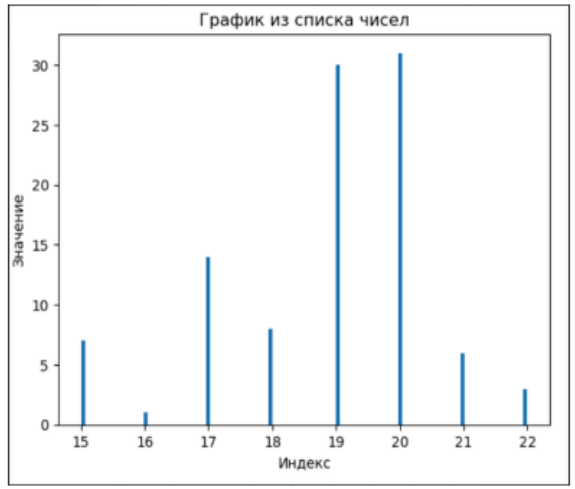

Рисунок 4. Количество ключей для промпта №1 с рассуждением

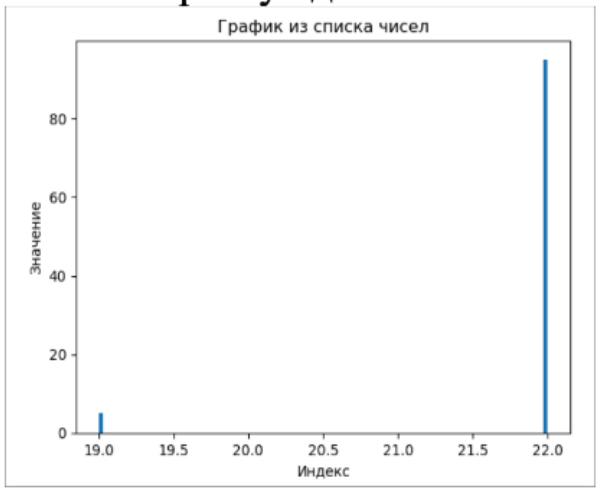

Рисунок 5. Количество ключей для промпта №2 без рассуждения

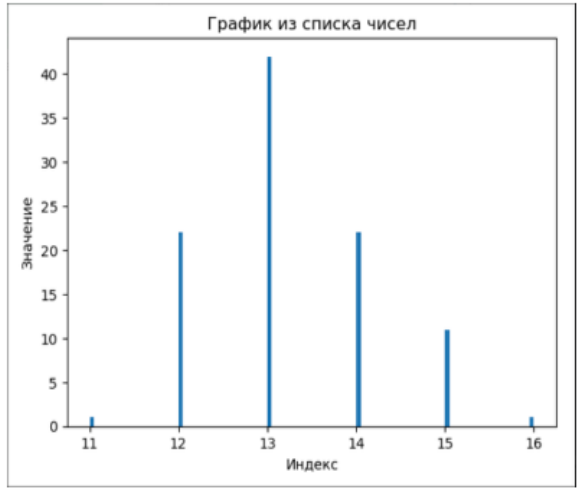

Рисунок 6. Количество ключей для промпта №2 с рассуждением

Вхождение в эталонный массив

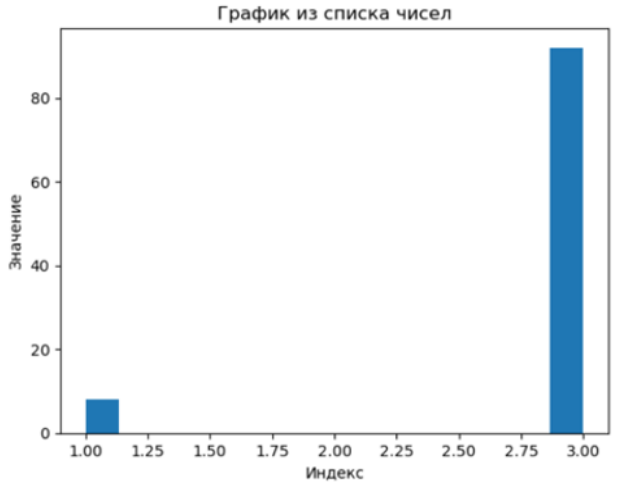

Рисунок 7. Вхождение в эталонный массив с промптом №1 без рассуждений

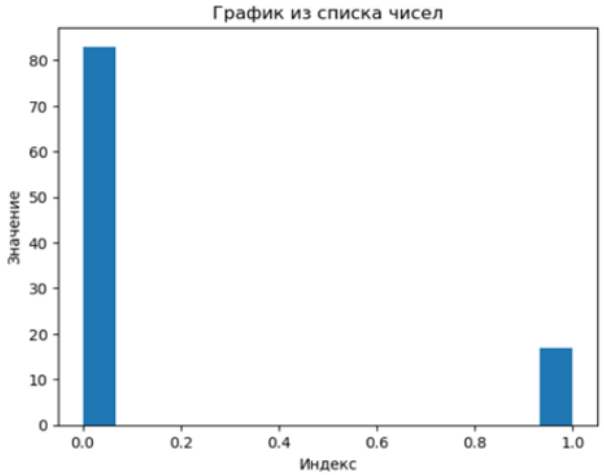

Рисунок 8. Вхождение в эталонный массив с промптом №1 с рассуждениями

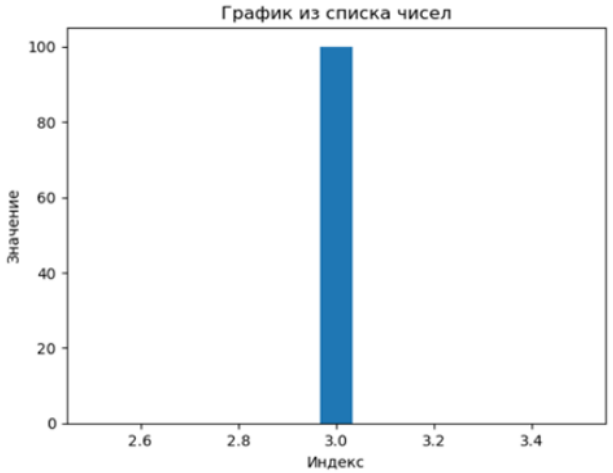

Рисунок 9. Вхождение в эталонный массив с промптом №2 без рассуждений

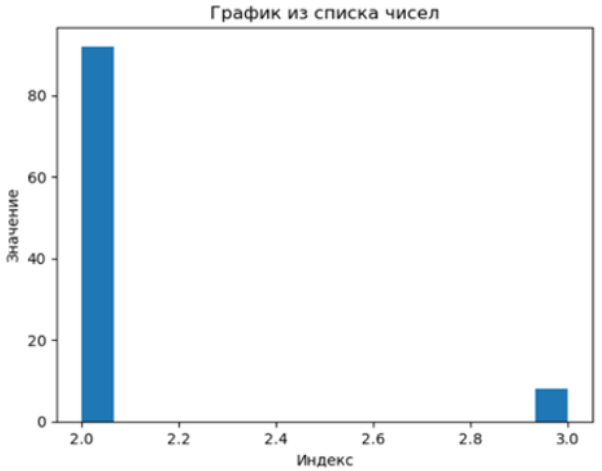

Рисунок 10. Вхождение в эталонный массив с промптом №2 с рассуждениями

t-SNE визуализация

Рисунок 11. t-SNE визуализация

t-SNE визуализация с Perplexity

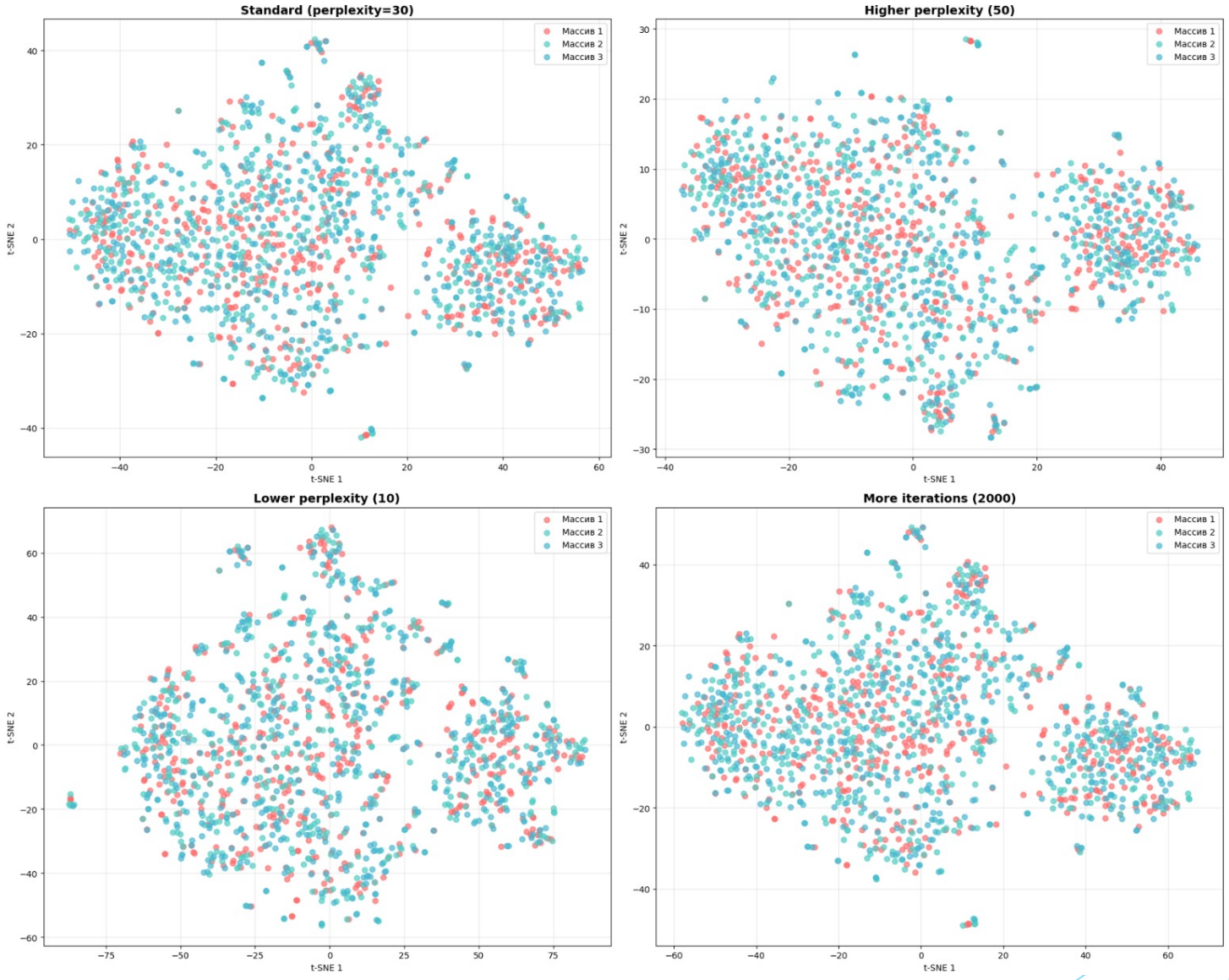

Рисунок 12. t-SNE визуализация эмбеддингов и текстов

Тепловая карта сходства усредненных эмбеддингов

Рисунок 13. Тепловая карта сходства усредненных эмбеддингов

Выводы

- Подтверждается Гипотеза 1, что для сокращения количества атрибутов логов возможно использовать на предварительном шаге эту же самую LLM, выполняя очистку «лишних» атрибутов.
- Подтверждается Гипотеза 2 – использование дополнительного агента с функцией размышления на этапе удаления ключей. При этом выбор моделью ключей достаточно стабильно попадает в эталонный массив ключей.
- Проверка представленных гипотез проводилась ограниченным набором методик работы с большими языковыми моделями, таких как промпт-инжиниринг и цепочки рассуждений. При этом особенностью работы являлась необходимость исключить такие методики, как дообучение модели и генерация с дополненным поиском. Таким образом адекватным является подход решения прикладных задач с минимальным использованием ресурсов.
- По результатам исследования остался следующий вопрос: насколько применимо использование одной LLM как для генерации, так и для фактчекинга. А если модели должны быть разными, то насколько и где находится эта граница.