**Протокол проведения приемочных испытаний**

г. Санкт-Петербург «15» ноября 2024 г.

# Объект предварительных испытаний:

Фреймворк быстрого прототипирования приложений на основе больших языковых моделей (БЯМ).

Обозначение – ProtoLLM.

# **Цель испытаний:**

* проверка соответствия фреймворка Proto.LLM требованиям xfcnyjuj технического задания (приложение Д) в составе Программы исследовательского центра в сфере искусственного интеллекта "Сильный искусственный интеллект в промышленности" (далее ИЦ СИИП).

**Участники предварительных испытаний:**

Калюжная А.В., старший научный сотрудник ИЦ СИИП, Университет ИТМО

Деева И.Ю., старший научный сотрудник ИЦ СИИП, Университет ИТМО

Ковальчук М.А., инженер ИЦ СИИП, Университет ИТМО

Филатова А.А., инженер ИЦ СИИП, Университет ИТМО

Погосова А.А., руководитель офиса операционной и проектной деятельности Ассоциации «Цифровые технологии в промышленности»

# Место проведения предварительных испытаний: Университет ИТМО, Санкт-Петербург, Биржевая линия, д. 14, ауд. 310а.

# Документы, предъявленные на предварительные испытания:

# Результаты испытаний

## Результаты комплексной проверки возможности интеграции БЯМ без необходимости внесения изменений в архитектуру приложения и интеграции с инструментами для оценки и сравнения эффективности различных конфигураций системы на основе БЯМ

Лиза -> Ярослав

**Способ проверки:** в соответствии с п. 6.1 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

Успешность работы подтверждается получением ответов от БЯМ и вычислением метрик, оценивающих корректность этих ответов. После успешного выполнения пайплайна оценки ответов умного ассистента в консоли отображается вывод, содержащий заданный вопрос и ответ от БЯМ с наивысшей вычисленной метрикой. Результат запуска показан на Рисунке 5.1.1.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рисунок 5.1.1 – Результат комплексной проверки (вывод в консоль)

Результат испытаний соответствует критерию п.6.1 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

## Результаты комплексной проверки возможности создания прототипов RAG систем с автоматизацией процессов индексации, обработки и загрузки данных в векторные БД, а также возможностью настройки процесса retrieval в системах на основе RAG

Настя

**Способ проверки:** в соответствии с п. 6.2 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

Результатом успешного завершения комплексной проверки применимости фреймворка для создания RAG системы с применением автоматизированных процессов индексации, обработки и загрузки данных в векторные БД, а также настройки процесса извлечения информации (retrieval) из баз данных, является получение ответа системы на пользовательский запрос с учетом контекста, полученного на этапе retrieval. Пример результата на рисунке 5.2.1 демонстрирует вывод информации о завершении процесса генерации и ответа на естественном языке в консоль.

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рисунок 5.2.1 – Результат работы системы генерации ответа на пользовательский запрос

Результат испытаний соответствует критерию п.6.2 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

## Результаты комплексной проверки возможности создания прототипов агентных систем с унифицированным способом для подключения внешних сервисов и приложений в виде плагинов

Алина:

**Способ проверки:** в соответствии с п. 6.3 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

В результате комплексной проверки агентного пайплайна в консоли должна отображаться информация об успешном вызове плагинов. В данном случае на рисунке 5.3.1 видно, что результатом обращения стала предложенный для запуска плагин для генерации лекарства по задаче Дислипидемия.

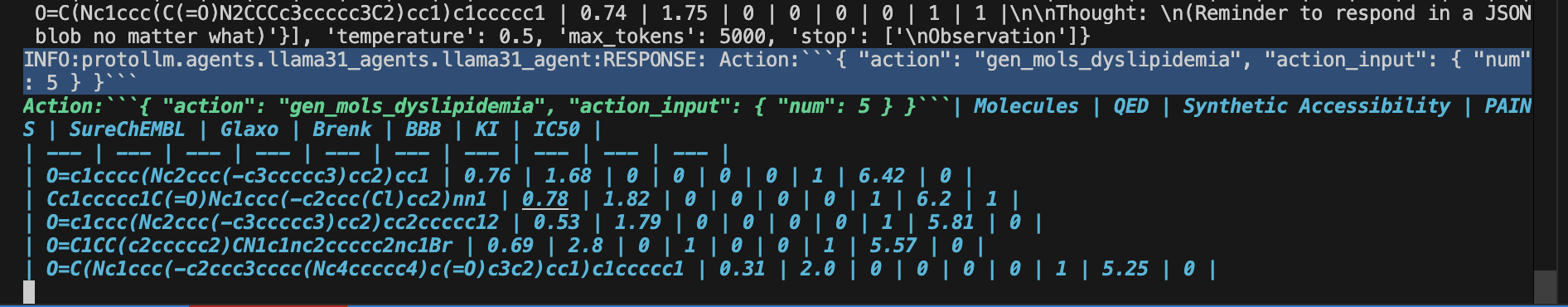


Рисунок 5.3.1 – Результат вызова внешнего плагина (вывод консоли) с помощью агентной системы

Результат испытаний соответствует критерию п.6.3 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

## Результаты проверки возможности создания, управления и использования ансамблей языковых моделей

**Способ проверки:** в соответствии с п. 6.4 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

По результатам подготовки пула моделей были получены следующие вывод консоли контейнеров балансировщика нагрузки (Рисунок 5.1.1), контейнеров сервера VLLM модели Qwen2.5-72b (Рисунок 5.1.2), идентифицирующие готовность моделей к обработке запросов.

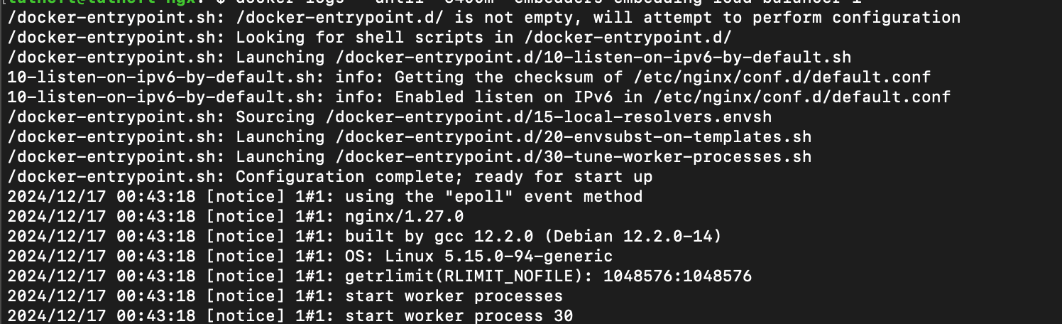
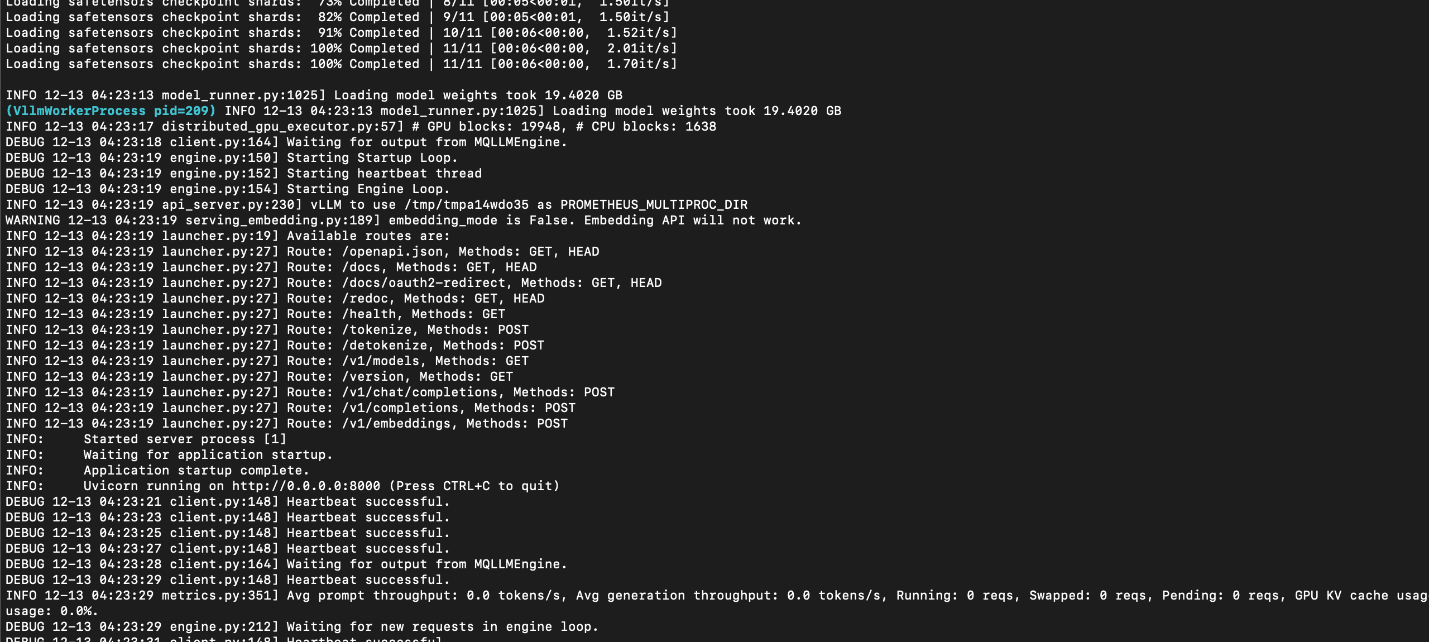
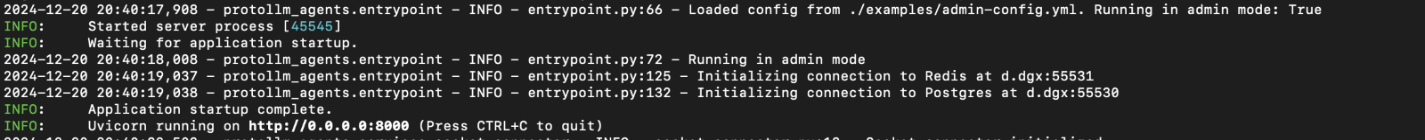


Рисунок 5.1.1 - Результат запуска балансировщика нагрузки NGINX.

Рисунок 5.1.2 - Пример результата запуска серверов сервинга моделей Qwen2.5-72b.

В результате инициализации агентов, пула модели, векторных хранилищ и запуска объекта Entrypoint с помощью конфигурационного файла, в логах сервиса можно видеть успешный старт сервера Uvicorn ([http://0.0.0.0:8080](http://0.0.0.0:8000)). Пример вывода консоли после старта сервиса приведен на Рисунке 5.1.3.

Рисунок 5.1.2 - Результат запуска компоненты модуля повышения эффективности БЯМ за счет использования ансамблей моделей и мультиагентных технологий в виде вывода консоли.

На OpenAPI эндпойнте сервиса FastAPI можно видеть информацию о добавленных агентах. Результат представлен на Рисунке 5.1.3.

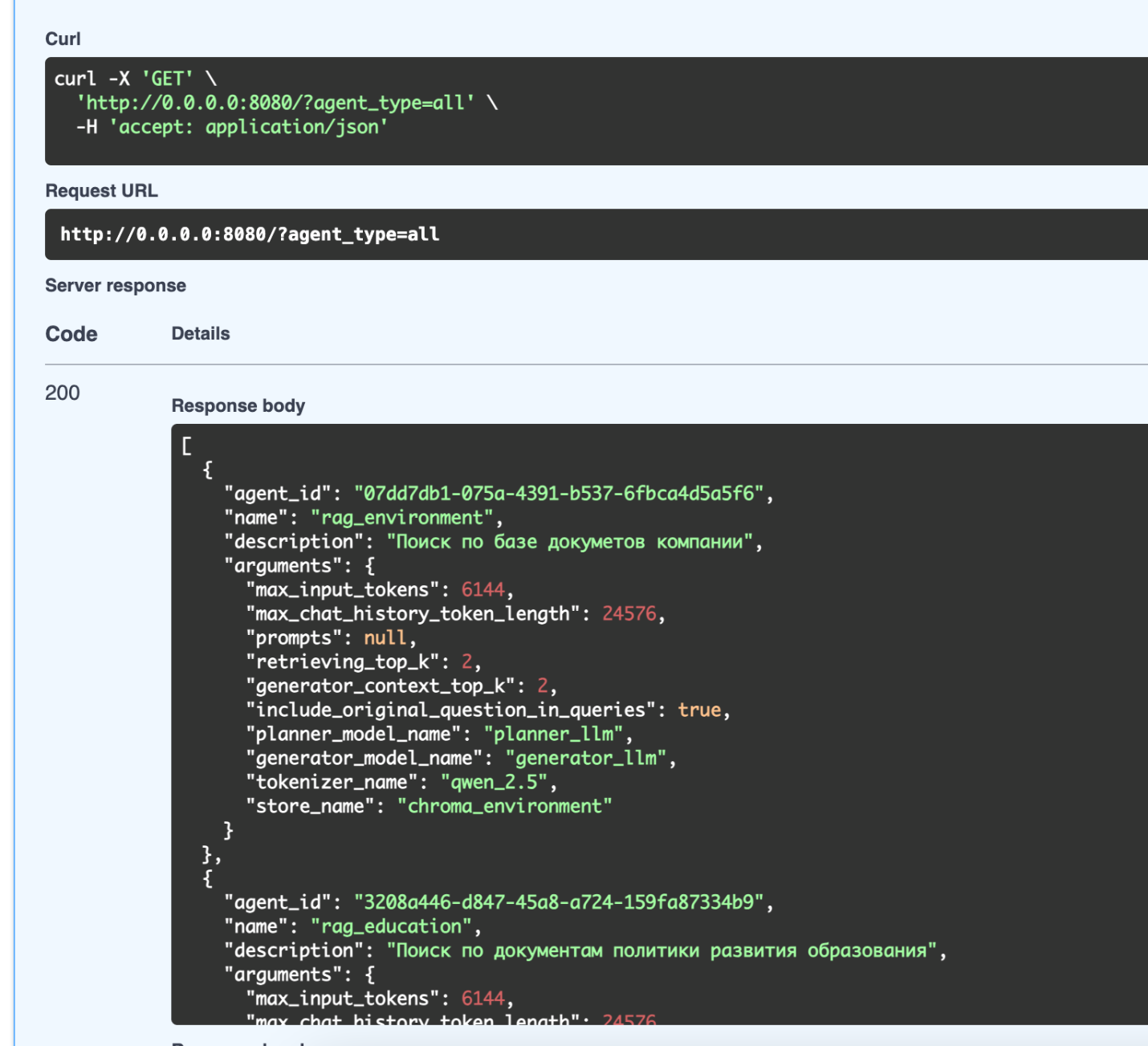


Рисунок 5.1.3 - Результат запуска компоненты модуля повышения эффективности БЯМ за счет использования ансамблей моделей и мультиагентных технологий в виде запроса листинга добавленных агентов на эндпойнте OpenAPI (сконфигурированные агенты).

Также можно видеть автоматически добавленных агентов агента-роутера и агента-синтеза. Результат листинга агентов (продолжение) представлен на Рисунке 5.1.4.

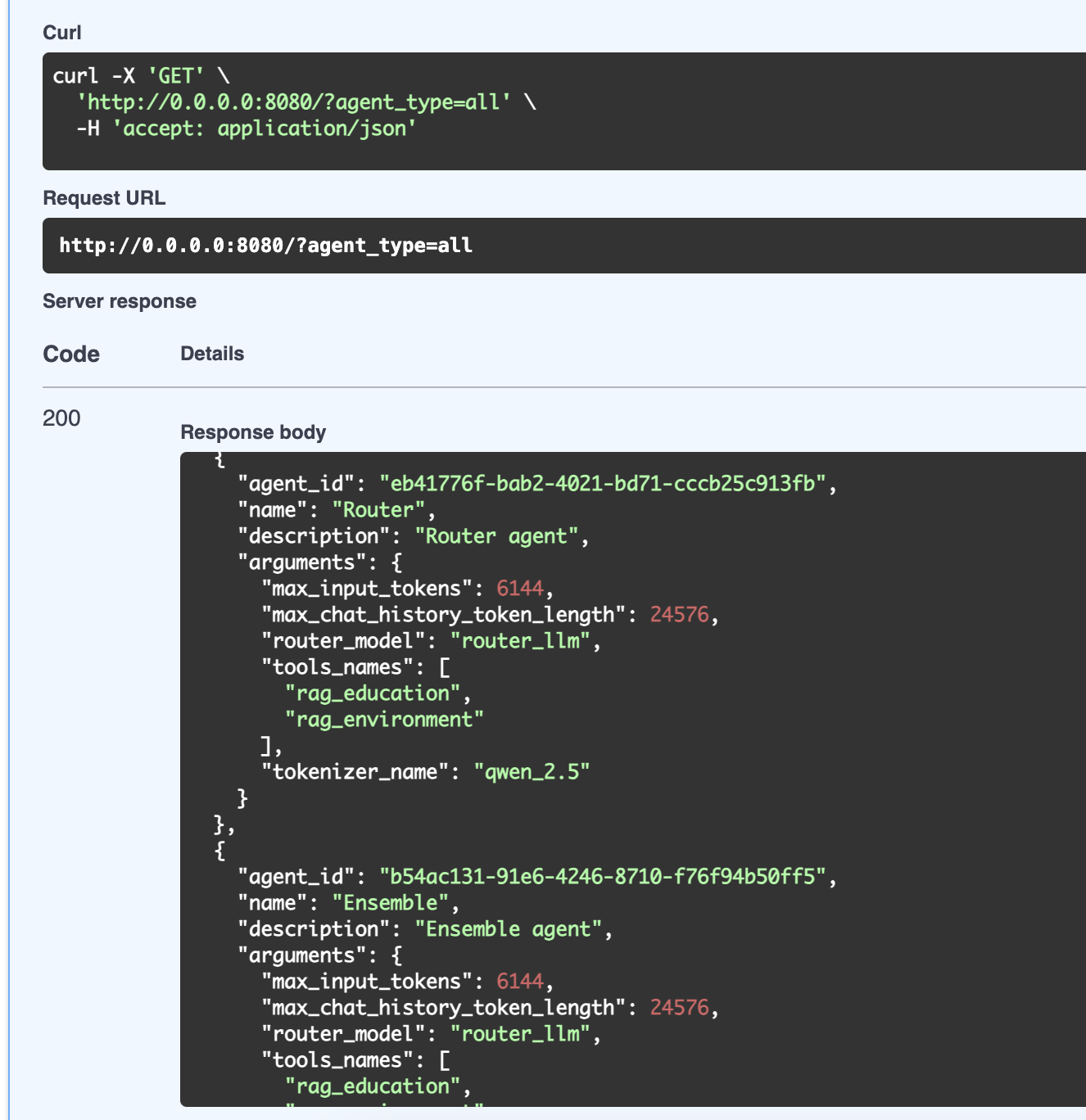


Рисунок 5.1.3 - Результат запуска компоненты модуля повышения эффективности БЯМ за счет использования ансамблей моделей и мультиагентных технологий в виде запроса листинга добавленных агентов на эндпойнте OpenAPI (автоматически добавленные агенты).

В результате запроса Postman к агенту с идентификатором `3208a446-d847-45a8-a724-159fa87334b9` (rag\_education), был получен ряд сообщений со стороны сервиса, в том числе набор документов-источников, а также стриминговый ответ - результат генерации.

Результат в виде извлеченных из векторного хранилища документов представлен на Рисунке 5.1.5. Полное событие представлено в Листинге 5.1.1.



Рисунок 5.1.5 - Результат сообщения события, полученного в качестве промежуточного результата запроса на Websocket API компоненты модуля повышения эффективности БЯМ за счет использования ансамблей моделей и мультиагентных технологий.

Листинг 5.1.1 - Сообщение события, полученного в качестве промежуточного результата запроса на Websocket API компоненты модуля повышения эффективности БЯМ за счет использования ансамблей моделей и мультиагентных технологий.

|  |
| --- |
| {  "event\_id": "0725657b-e33c-4cc1-9be6-13ab0e70625c",  "agent\_id": "3208a446-d847-45a8-a724-159fa87334b9",  "description": null,  "is\_eos": false,  "name": "retrieval",  "result": [  {  "id": null,  "metadata": {  "chapter": "ПОСТАНОВЛЕНИЕ",  "file\_name": "education.docx",  "is\_heading\_extracting\_correct": false,  "original\_file\_name": "education.docx",  "original\_source": "/docs/education.docx",  "source": "/docs/education.docx"  },  "page\_content": "<tr><td>5 </td><td>Задачи государственной программы </td><td>Повышение качества и доступности образования всех уровнейдля жителей Санкт-Петербурга;повышение уровня квалификации, условий труда и уровня заработной платы руководящих, педагогических, иных категорий работников системы образования Санкт-Петербурга;повышение уровня материально-технической базы и развитие инфраструктуры системы образования Санкт-Петербурга;обеспечение соответствия профессиональных образовательных программ потребностям экономики Санкт-Петербурга;повышение роли Санкт-Петербурга как всероссийского и международного научно-образовательного центра </td></tr><tr><td>6 </td><td>Целевые показатели государственной программы </td><td>Доступность дошкольного образования для детей в возрасте от 3 до 7 лет;доступность дошкольного образования для детей в возрасте от 2 месяцев до 3 лет;удельный вес численности выпускников профессиональных образовательных организаций очной формы обучения, трудоустроившихся в течение первого года после окончания обучения, в общей численности выпускников образовательных организаций профессионального образования очной формы обучения </td></tr>\n<tr><td>7 </td><td>Основания разработки государственной программы </td><td>;, утвержденная",  "type": "Document"  },  {  "id": null,  "metadata": {  "chapter": "ПОСТАНОВЛЕНИЕ",  "file\_name": "education.docx",  "is\_heading\_extracting\_correct": false,  "original\_file\_name": "education.docx",  "original\_source": "/docs/education.docx",  "source": "/docs/education.docx"  },  "page\_content": "<tr><td>2 </td><td>Соисполнители государственной программы </td><td>Администрации районов Санкт-Петербурга;Комитет имущественных отношений Санкт-Петербурга;Комитет по здравоохранению;Комитет по культуре Санкт-Петербурга;Комитет по науке и высшей школе;Комитет по социальной политике Санкт-Петербурга;Комитет по строительству;Комитет по физической культуре и спорту </td></tr><tr><td>3 </td><td>Участники государственной программы </td><td>- </td></tr>\n<tr><td>4 </td><td>Цели государственной программы </td><td>Обеспечение высокого качества и доступности образования для всех слоев населения в интересах социально-экономического развития Санкт-Петербурга",  "type": "Document"  },  {  "id": null,  "metadata": {  "chapter": "ПОСТАНОВЛЕНИЕ",  "file\_name": "education.docx",  "is\_heading\_extracting\_correct": false,  "original\_file\_name": "education.docx",  "original\_source": "/docs/education.docx",  "source": "/docs/education.docx"  },  "page\_content": "\n<tr><td>10</td><td>Ожидаемые результаты реализации государственной программы </td><td>Создание в Санкт-Петербурге инфраструктуры поддержки раннего развития детей (0-3 года);отсутствие очереди на зачисление детей в возрастеот 3 до 7 лет в дошкольные образовательные организации Санкт-Петербурга;создание во всех общеобразовательных организациях Санкт-Петербурга условий, соответствующих требованиям федеральных государственных образовательных стандартов;повышение результатов обучающихся общеобразовательных организаций Санкт-Петербурга по итогам международных сопоставительных исследований качества общего образования (PIRLS, TIMSS и PISA);создание условий для получения жителями Санкт-Петербурга профессиональной подготовки, переподготовки и повышения квалификации;охват детей в возрасте от 5 до 18 лет программами дополнительного образования на уровне не менее 85 процентов;повышение привлекательности педагогической профессии и уровня квалификации преподавательских кадров Санкт-Петербурга",  "type": "Document"  }  ]  } |

Результат генерации ответа на запрос представлен на Рисунке 5.1.6. Промежуточный результат генерации ответа, а также конечный результат, представлены в Листинге 5.1.2.

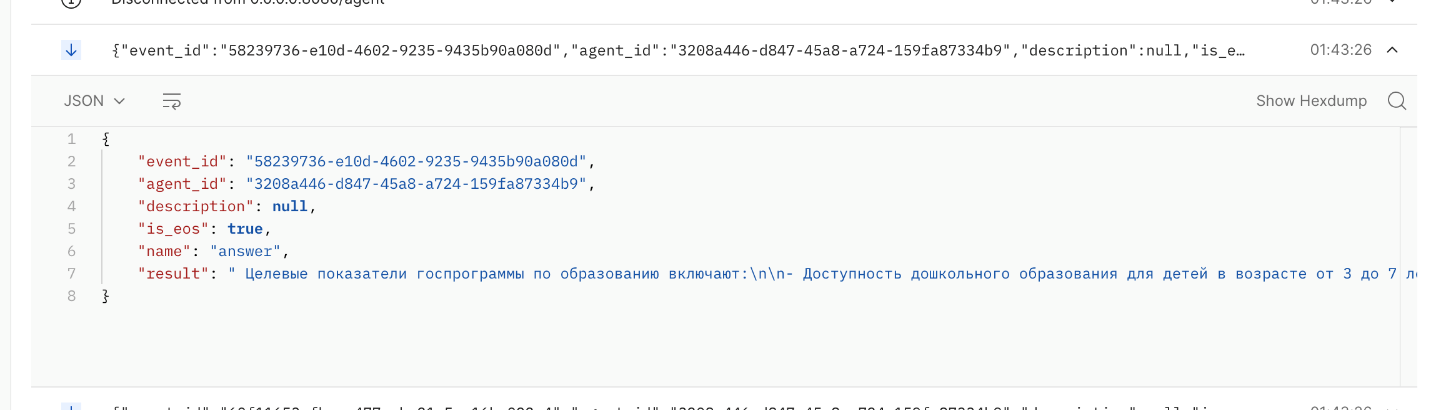
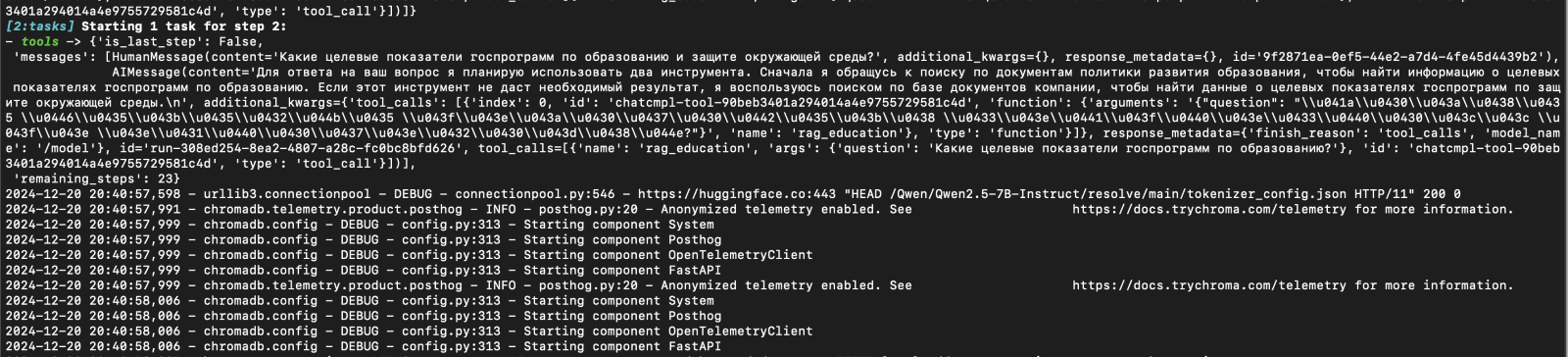


Рисунок 5.1.6 - Результат сообщения события, полученного в качестве промежуточного результата запроса на Websocket API компоненты модуля повышения эффективности БЯМ за счет использования ансамблей моделей и мультиагентных технологий.

В результате запросов на эндпойнты `/router` и `/ensemble` были получены как промежуточные сообщения с извлеченными документами, так и сообщения генерации ответов (промежуточной и финальной). Пример вывода консоли во время запроса к агенту-роутеру представлен на Рисунке 5.1.8, можно видеть, что агент-роутер вызвал инструмент, полученный из агента `rag\_education`.

  
Рисунок 5.1.8 - Вывод консоли, являющийся результатом вызова агентом-роутером инструментов, составленных на базе существующих в системе агентов.

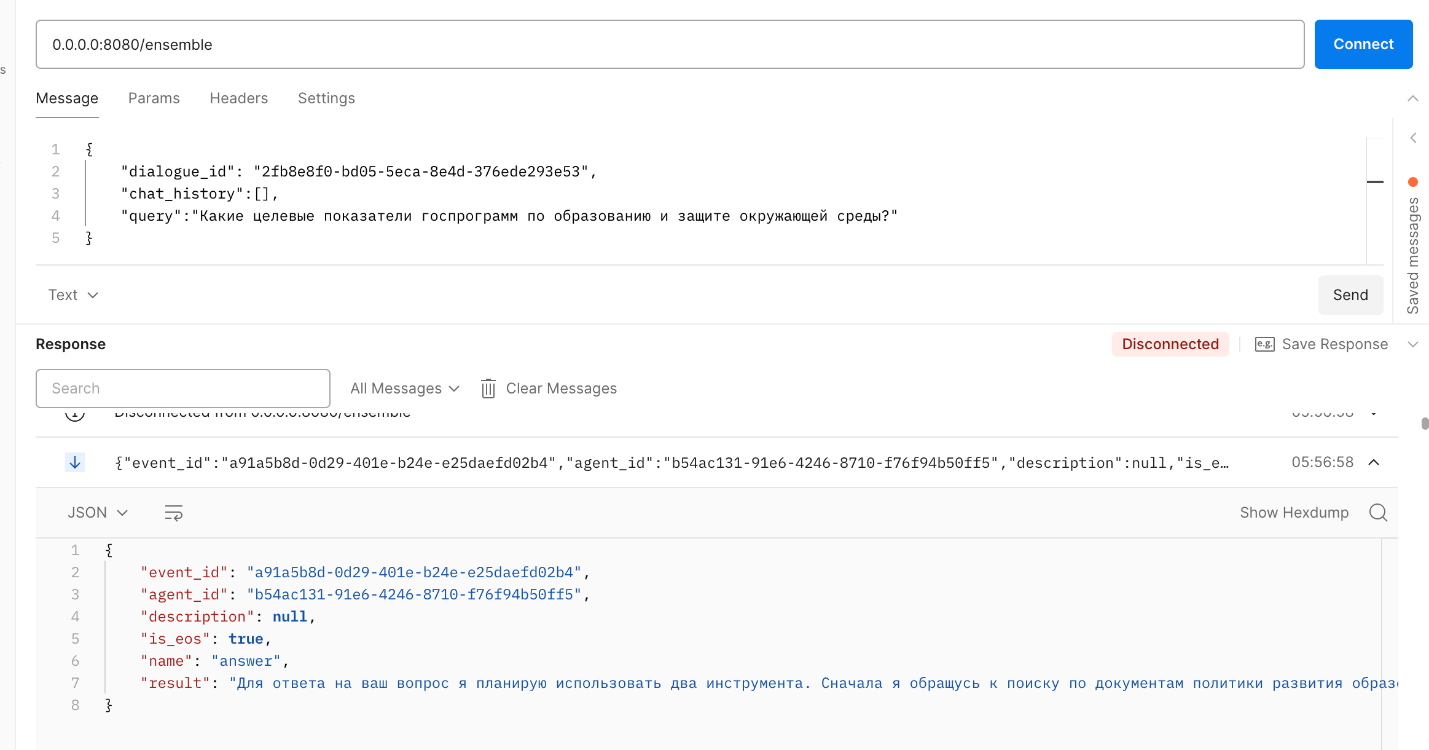
На Рисунке 5.1.9 представлен интерфейс Postman, в котором развернуто последнее сообщение системы, полученное от агента ансамблирования (is\_eos=False).

Рисунок 5.1.9 - Результат полученного события, содержащего результат генерации ответа на запрос на Websocket API компоненты модуля повышения эффективности БЯМ за счет использования ансамблей моделей и мультиагентных технологий.

В ходе испытаний была заполнена таблица результатов. Структура таблицы приведена на рисунке 5.1.10

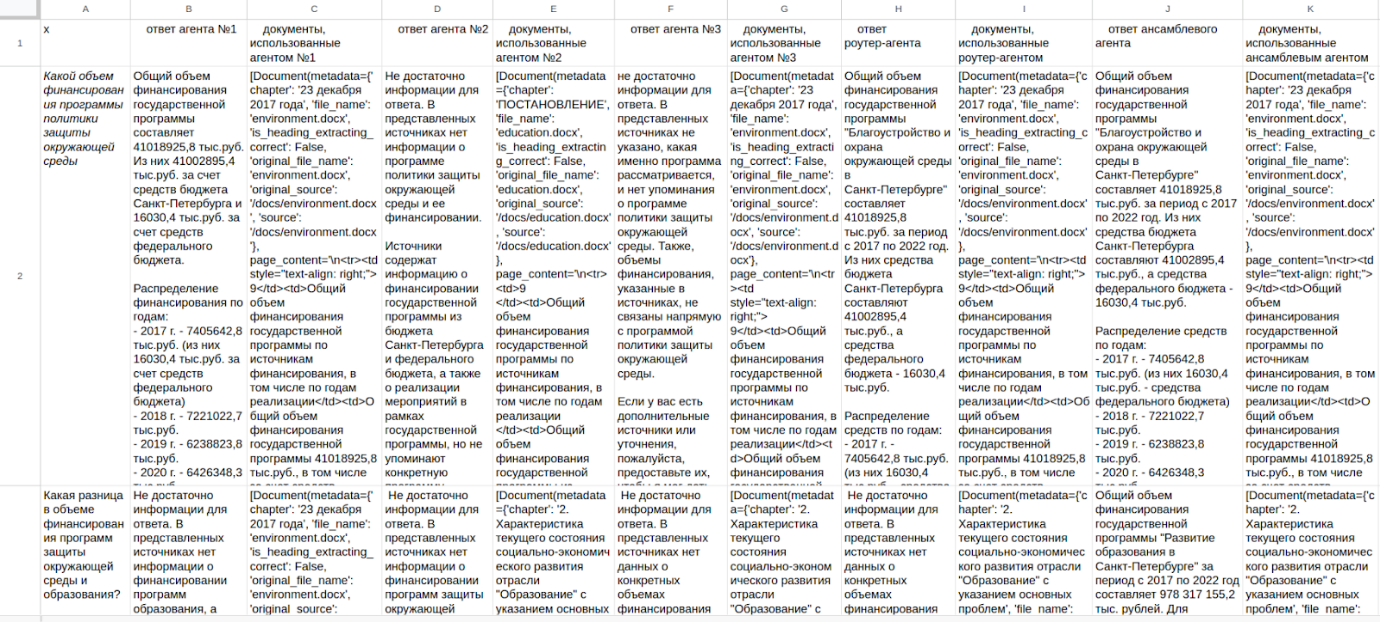


Рисунок 5.1.10 - Структура таблицы результатов генерации.

При проверке было определено, что на каждый вопрос был дан ответ соответствующим агентом, каждая ячейка с выделенными документами не пуста.

Результат испытаний соответствует критерию п.6.4 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

## Результат проверки возможности генерации сложных синтетических примеров, пригодных для дообучения БЯМ

Маша:

**Способ проверки:** в соответствии с п. 6.5 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

…

Результат испытаний соответствует критерию п.6.5 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

## Результат комплексной проверки комплектности программных модулей и программной документации

Аня:

**Способ проверки:** в соответствии с п. 6.6 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

…

Результат испытаний соответствует критерию п.6.6 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

## Результат комплексной проверки наличия шаблонов для типовых систем с RAG и агентами

Коля:

**Способ проверки:** в соответствии с п. 6.7 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

В результате проверки подтверждено наличие необходимых шаблонов. Шаблоны запросов располагаются в каталоге prompt\_templates (12 шаблонов для различных сценариев использования моделей, 5 шаблонов для генерации синтетических данных). Скриншот каталога, содержащего файлы шаблонов, показан на рисунке 5.7.1.

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Рисунок 5.7.1 – Файлы с шаблонами запросов.

Файла с шаблонами кода распложены в каталоге code\_templates, всего представлено 3 шаблона, покрывающих все требуемые сценарии (агентная БЯМ, подключение плагинов, RAG). Скриншот каталога, содержащего файлы шаблонов, показан на рисунке 5.7.2.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 5.7.2 – Файлы с шаблонами кода.

Результат испытаний соответствует критерию п.6.7 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

## Результат проверки наличия обработки сбоев и исключений

Коля:

**Способ проверки:** в соответствии с п. 6.8 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

В результате проверки подтверждено корректная обработка исключений на примере ситуации с передачей некорректного значения входного параметра пути к модели. Пример ошибки, выводимой в консоли, показан на рисунке 5.8.1.

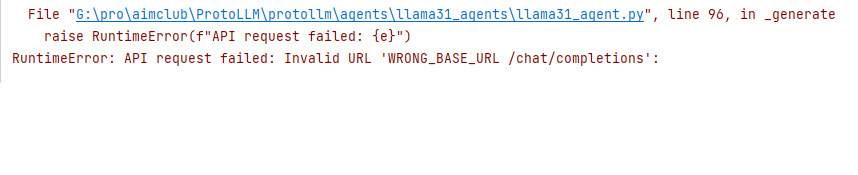


Рисунок 5.8.1 – Вывод в консоль с обработкой ошибок

Результат испытаний соответствует критерию п.6.8 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

## Результат проверки наличия сведений о порядке обращения к разработчикам при выявлении сбоев и ошибок при работе библиотеки

Коля

**Способ проверки:** в соответствии с п. 6.9 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

Проверка подтвердила наличие информации с описанием порядка обращения к разработчикам. Скриншот из англоязычной версии README показан на Рисунке 5.9.1.

Изображение выглядит как текст, Шрифт, снимок экрана, линия

Автоматически созданное описание

Рисунок 5.9.1 – Скриншот раздела с контактами разработчиков

Результат испытаний соответствует критерию п.6.9 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

## Результат проверки наличия возможности предоставления API для программного управления всеми основными функциями системы

Юра

**Способ проверки:** в соответствии с п. 6.10 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

…

Результат испытаний соответствует критерию п.6.10 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

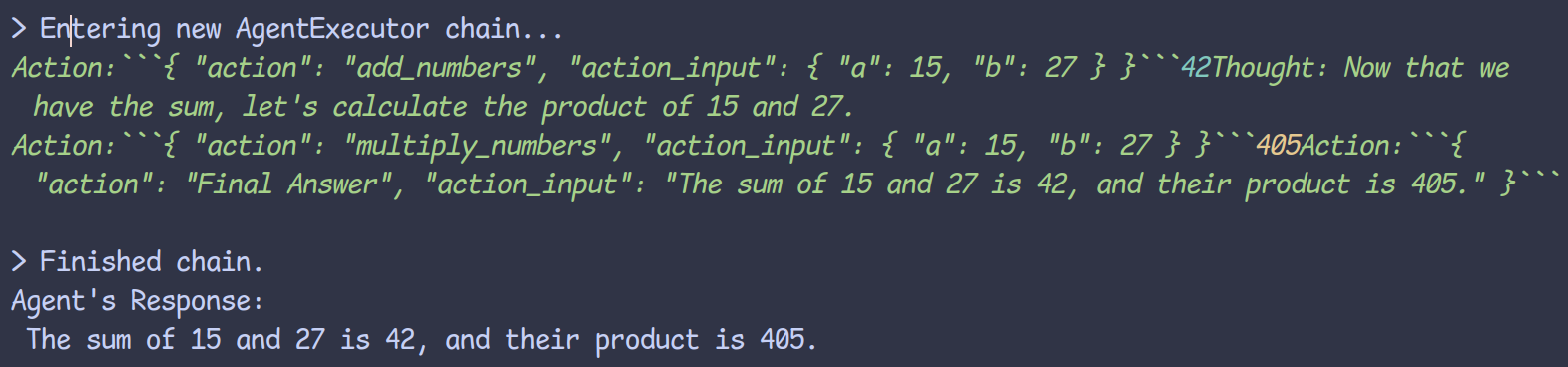
## Результаты проверки поддержки механизмов для обеспечения воспроизводимости экспериментов с различными конфигурациями БЯМ и плагинов.

Юра

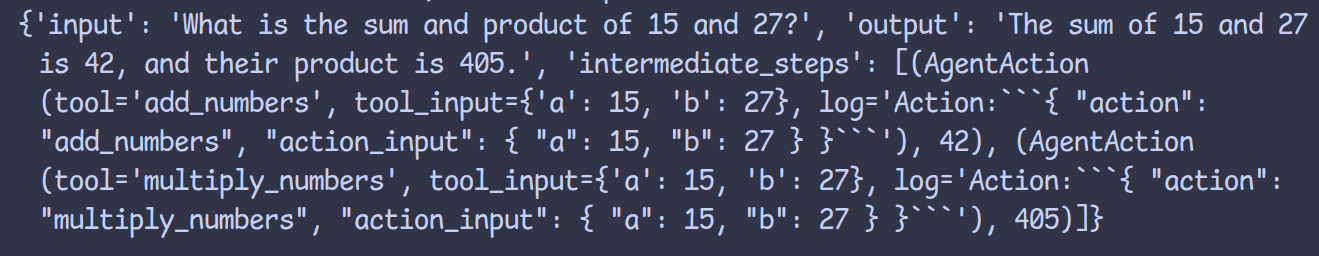
**Способ проверки:** в соответствии с п. 6.11 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

При запуске сценария БЯМ с плагинами, в режиме реального времени в терминале будет отображаться статус работы БЯМ над задачей как показано на рисунке 5.10.1.

Рисунок 5.10.1 – Скриншот статуса работы БЯМ с плагинами в терминале

После успешного завершения, возвращается словарь как на рисунке 5.10.2.

Рисунок 5.10.2 – Скриншот итогового словаря, содержащего ответ БЯМ и промежуточные шаги

Результат испытаний соответствует критерию п.6.11 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

## Результат комплексной проверки быстродействия и точности конфигураций системы на основе БЯМ

Аня+Коля+Лиза (я добавлю описание из статьи, Коля или Лиза добавят подходящие кусочки кода из репо к статье):

**Способ проверки:** в соответствии с п. 6.12 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

…

Результат испытаний соответствует критерию п.6.12 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

## Проверка наличия тестов

Коля

**Способ проверки:** в соответствии с п. 6.13 программы и методики испытаний фреймворка Proto.LLM RU.СНАБ.00855-01 51 01.

**Результат проверки:**

Проверка наличия модульных и интеграционных тестов подтвердила их наличие в каталогах tests, protollm\_tools/llm-api/tests, protollm\_tools/sdk/tests. Скриншот каталога с интеграционными тестами модуля llm-api приведен на рисунке 5.13.1. Отчет об их успешном прохождении на рисунке 5.13.2.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 5.13.1 – Скриншот каталога с интеграционными тестами модуля llm-api

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

Рисунок 5.13.2 – Отметка о успешном прохождении теста

# Замечания и рекомендации:

Замечаний нет

# Выводы

**Испытания проводили:**

|  |  |
| --- | --- |
| Старший научный сотрудник ИЦ СИИП Университета ИТМО, к.т.н. | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.В. Калюжная |
| Старший научный сотрудник ИЦ СИИП Университета ИТМО | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ И.Ю. Деева |
| Инженер ИЦ СИИП Университета ИТМО | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ М.А. Ковальчук |
| Инженер ИЦ СИИП Университета ИТМО | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.А. Филатова |
| Руководитель офиса операционной и проектной деятельности, Ассоциации «Искусственный интеллект в промышленности» | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.А. Погосова |