УДК 004.82:519.768

ИНТЕГРАЦИЯ ОНТОЛОГИЙ И БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ОТВЕТА НА НАУКОМЕТРИЧЕСКИЕ ВОПРОСЫ НА ПРИМЕРЕ ТЕОРИИ УПРАВЛЕНИЯ

Д. А. Губанов*, В. А. Сергеев**

*,**Институт проблем управления им. В. А. Трапезникова РАН, г. Москва

*✉ dmitry.a.g@gmail.com, **✉ sergeev.bureau@gmail.com

Аннотация. Рассматривается задача автоматического формирования ответов на сложные наукометрические вопросы к базам знаний, заданные на естественном языке. Актуальность исследования обусловлена ограничениями современных больших языковых моделей (англ. *Large Language Model*, LLM), которые, несмотря на высокую степень «понимания» вопросов, склонны генерировать неточные ответы и не всегда обладают актуальными сведениями в специализированных предметных областях. В то же время графы знаний обеспечивают точность и актуальность информации, однако требуют знания формальных языков запросов. Предложено решение на основе гибридной архитектуры, в которой LLM выполняет функцию интеллектуального интерфейса к онтологической базе знаний, преобразуя вопросы на естественном языке в корректные SPARQL-запросы, результаты выполнения которых возвращаются пользователю. Для решения задачи составлен специализированный корпус данных для обучения и тестирования NL-to-SPARQL моделей в области теории управления. Подход реализован на основе онтологии научной деятельности в области теории управления и апробирован на сформированном корпусе вопросов. Интеграция LLM с онтологической базой знаний позволила добиться высокой точности ответов (около 99 %), что подтверждает перспективность предложенного подхода.

Ключевые слова: наукометрия, графовые базы знаний, большие языковые модели.

ВВЕДЕНИЕ

В последние годы большие языковые модели (см., например, работы [1–5]) показывают впечатляющие результаты в задачах обработки естественного языка и тем самым открывают новые возможности для интеллектуального поиска в базах данных. Способность трансформерных архитектур улавливать семантический контекст запроса позволила перейти от поиска по ключевым словам к поиску по смыслу, однако применение таких моделей в специализированных предметных областях сопряжено с серьезными ограничениями. При решении задач, требующих точности и актуальности (таких, как наукометрический анализ) модели склонны к «галлюцинациям» (генерации правдоподобных, но ложных ответов) и не способны оперировать данными, появившимися после завершения их обучения [6–8].

Альтернативным подходом является применение структурированных баз знаний и технологий Linked Data. Онтологии обеспечивают строгое толкование терминов, а графы знаний (англ. *Knowledge Graphs*, KG) служат достоверным источником фактов. Доступ к данным графа знаний осуществляется при помощи таких формальных языков запросов, как SPARQL. Такой способ поиска гарантирует интерпретируемость и воспроизводимость результатов, что критически важно для анализа научной деятельности. Тем не менее, использование SPARQL создает высокий барьер входа: составление корректного запроса требует специальных знаний и навыков, а существующие графические интерфейсы ограничивают пользователя набором типовых фильтров.

Задача автоматической трансляции естественного языка в формальные запросы (NL-to-SPARQL) призвана устранить этот барьер. Несмотря на прогресс в этой области [9], существующие решения сталкиваются с проблемами при обработке комплексных аналитических запросов. В отличие от простых «фактоидных» вопросов (например, «*Кто автор статьи X?*»), аналитические вопросы в наукометрии часто требуют многоэтапной обработки, агрегирования, сравнения количественных показателей и фильтрации по сложным условиям (например, «*Найти наиболее активных по теме A авторов* (*за последние 5 лет*) *из организации B, которые прямо или косвенно* (*по цепочке*) *процитировали ученого C*»). Стандартные подходы часто генерируют синтаксически неверный код или ошибаются в логике связей между сущностями.

Научная аналитика и наукометрия представляют собой прикладные области, где выгоды от объединения больших языковых моделей с семантическими технологиями особенно велики. Объем научной информации – публикаций, данных об исследователях, цитирований и т. д. – неуклонно растет, и для его осмысления требуются методы интеллектуального анализа. Платформа Linked Data предоставляет инструменты для интеграции разнородных сведений о научной деятельности в единый граф знаний с формализацией понятий (например, исследователь, статья, журнал, организация, тема исследования) и связей между ними (авторство, цитирование, принадлежность к организации, развитие темы и др.). Уже существуют крупные открытые графы знаний такого рода, например, Microsoft Academic Knowledge Graph (MAKG) или SemOpenAlex (преемник MAKG), содержащие миллиарды триплетов о публикациях, авторах и организациях [10, 11].

В данной работе предлагается решение описанной выше проблемы на примере такой предметной области, как теория управления. Представлен подход к интеграции больших языковых моделей с онтологической базой знаний ИСАНД (Информационная система анализа научной деятельности) [12]. Предлагаемый метод использует LLM как интеллектуальный интерфейс, преобразующий вопросы на естественном языке в валидные SPARQL-запросы к базе знаний в конкретной научной области.

Научный вклад работы заключается в следующем. Разработана архитектура гибридной QA-системы, которая адаптирована для обработки сложных аналитических запросов в узкоспециализированных научных областях, в отличие от аналогов, ориентированных на открытые и широкие предметные области (Wikidata) и фактоидные вопросы. Создан и описан русскоязычный набор данных (датасет) по теории управления, восполняющий дефицит ресурсов для обучения NL-to-SPARQL-моделей в специализированных предметных областях. Экспериментально подтверждена эффективность дообучения (*fine-tuning*) LLM для задачи генерации SPARQL-кода для сложной схемы данных.

Структура статьи следующая. В § 1 описаны применяемые методы и архитектура гибридной системы, в § 2 приведено описание данных – онтологической базы знаний и сформированного набора запросов, в § 3 представлены результаты экспериментов и их анализ. Наконец, в заключении приводятся выводы и направления будущих работ.

1. МЕТОДЫ

Ранние исследования в области преобразования вопросов на естественном языке в SPARQL-запросы базировались преимущественно на методах семантического парсинга, использовании шаблонов и правил. Такие подходы показали приемлемое качество при работе с простыми, прямолинейными запросами, однако обладали существенными ограничениями в обработке сложных запросов, включающих вложенные конструкции или агрегирующие функции. Первые системы, такие как PowerAqua [13] или Aqqu [14], использовали наборы шаблонов для отображения языковых триплетов на естественном языке (ЕЯ, англ. *natural language*, NL) в онтологические триплеты графов знаний, что существенно ограничивало их применимость только простыми по структуре вопросами. С появлением крупных корпусов пар «вопрос NL – запрос SPARQL», таких как QALD (*Question Answering over Linked Data*) – см., например, статью [9] – стало возможным применение нейросетевых моделей. Генерация SPARQL-запросов стала рассматриваться как задача машинного перевода, в которой вопрос, сформулированный на естественном языке, выступает исходным текстом, а SPARQL-запрос – целевым текстом. Например, в работе [15] был предложен подход на основе нейронных сетей на базе LSTM (*long short-term memory*). Однако, несмотря на значительный прогресс, обработка сложных многокомпонентных запросов все еще представляет проблему.

В настоящее время в качестве перспективного направления рассматривается интеграция LLM с графовыми базами знаний для построения продвинутых QA-систем. Можно выделить несколько ос-

новных схем интеграции больших языковых моделей с онтологиями научной деятельности на основе графовых баз данных.

1. **Большая языковая модель как интерфейс к графу знаний**. Модель генерирует SPARQL-запрос (или иной формальный запрос) на основе вопроса пользователя, этот запрос затем выполняется на графе знаний, и полученный результат возвращается пользователю. Подход позволяет скрыть от пользователя сложность языка запросов, однако требует гарантированной корректности сформированного LLM-запроса (см. например, публикацию [16]).

2. **Граф знаний как источник знаний для большой языковой модели**. Прежде чем сформировать ответ, модель извлекает из графа релевантные факты (например, находит связанные триплеты по ключевым сущностям вопроса) и включает их в свой входной контекст. Такой вариант реализует стратегию Retrieval-augmented generation (RAG), при которой внешняя база знаний используется для формирования подсказки (*prompt*), что повышает фактическую точность ответа (см., например, работы [17, 18]).

3. **Граф знаний для верификации и доработки ответов большой языковой модели**. Модель сперва генерирует черновой ответ или список гипотез (возможно, с промежуточными рассуждениями), после чего подключается база знаний – система проверяет и уточняет результаты, например, отфильтровывая или ранжируя сгенерированные ответы на основе их подтверждаемости триплетами либо исправляя фактические ошибки (см., например, работу [19]).

Подходы, указанные в п. 2 и 3, могут выполняться итеративно: модель поэтапно, направляет запросы к базе знаний (например, посредством серии SPARQL-запросов) в духе цепочки рассуждений (*chain-of-thought*), постепенно собирая информацию, необходимую для ответа. Мировая тенденция такова, что современные системы вопросов – ответов все чаще строятся на подобных гибридных архитектурах (большая языковая модель и граф знаний), способных отвечать на сложные запросы, сохраняя при этом высокую естественность диалога с пользователем.

Для того, чтобы большая языковая модель правильно преобразовывала вопросы на ЕЯ в SPARQL-запросы (избегая галлюцинации схем, например выдумывания несуществующих предикатов), ее необходимо обучить. Это можно сделать, применяя подсказки (промпты) во время вывода (*in-context learning*) или применяя дообучение модели (*fine-tuning*) до вывода. Выбор в пользу дообучения обусловлен тремя факторами. Во-первых, сложностью схемы данных – полное описание онтологии в промпте исчерпает контекст локальной модели. Дообучение позволяет зафиксировать знания о структуре графа в весах модели. Во-вторых, надежностью синтаксиса. Дообучение значительно снижает вероятность генерации некорректного SPARQL-кода по сравнению с методами промпт-инжиниринга. В-третьих, эффективностью вывода. Использование в реально работающей системе дообученной модели меньшего размера (например, 7B/8B) экономически выгоднее и вычислительно эффективнее, чем использование огромных промптов с примерами в каждом запросе.

Для дообучения современной LLM могут понадобиться значительные вычислительные ресурсы, поэтому представляется целесообразным применение методов PEFT (*Parameter-Efficient Fine-Tuning*), которые снижают требования к используемому для вычислений оборудованию (см., например, публикацию [20]). В частности, разумным представляется использование методов LoRA (*Low-Rank Adaptation*) и QLoRA (*Quantized Low-Rank Adaptation*) [21, 22]. Метод QLoRA, по сравнению с LoRA, позволяет еще больше снизить потребление памяти, что важно для очень больших моделей. Эти методы позволяют добавить к предобученной модели дополнительные слои обучения с меньшим числом параметров относительно слоев исходной модели. Добавленные слои функционально выделяют в отдельный блок, называемый адаптером. Адаптеры обучаются под конкретную задачу, не изменяя веса исходной модели. Обучение и работа с адаптерами производились на двух видеокартах A30 с суммарным объёмом памяти GPU 48 Гб под управлением операционной системы openSUSE Leap 15.6.

В данной работе использовалась первая из перечисленных выше схем интеграции больших языковых моделей с онтологиями научной деятельности на основе графовых баз данных. Архитектура системы представлена на рис. 1. Основные функциональные блоки системы: блок LLM и блок KG (блок графа знаний). Работа системы происходит следующим образом. Пользователь формулирует запрос к системе. На основе полученного от пользователя запроса система формирует промпт для большой языковой модели. Далее промпт преобразуется с помощью большой языковой модели в SPARQL-последовательность. При обработке SPARQL-последовательности производится отсечение лишнего кода, сгенерированного LLM, комментируются служебные строки, а также добавля-

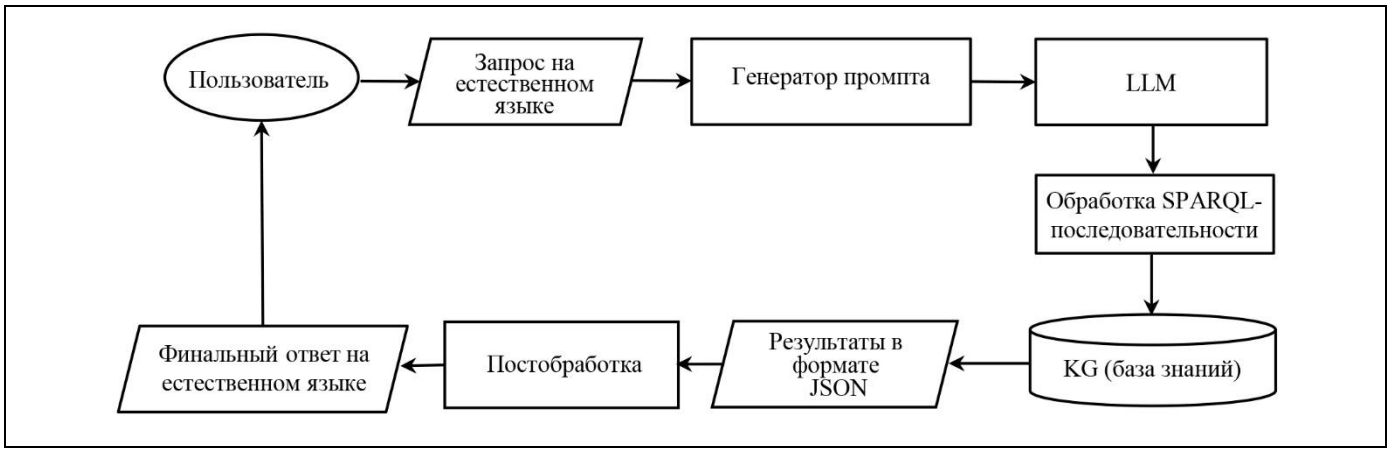

Рис. 1. Архитектура системы

ются префиксы для работы с базой данных. Модуль выполнения запросов направляет SPARQL-последовательность к базе данных и обрабатывает сетевые ошибки. С помощью полученной SPARQL-последовательности искомая информация в JSON-формате извлекается из графа знаний и после обработки возвращается пользователю.

2. ДАННЫЕ

Авторами сформирован уникальный набор данных, состоящий из пар вопрос – ответ по теории управления, включающий реальные информационные запросы экспертного сообщества. В отличие от существующих бенчмарков, этот набор данных сфокусирован на области теории управления и содержит качественно аннотированные пары вопрос – ответ. Пара состоит из вопроса пользователя к системе и ответа, являющегося SPARQL-запросом, корректного с точки зрения используемой онтологии ИСАНД. Все пары проверены специалистами. Такой подход к сбору данных обеспечивает высокую достоверность и релевантность тестовых примеров, а также отражает нюансы русскоязычной терминологии теории управления (что часто упускается в глобальных англоязычных наборах данных). При формировании набора данных первоначально экспертами были составлены 126 вопросно-ответных пар, корректных с точки зрения извлечения информации из рассматриваемого графа знаний.

В общем количестве примеров представлены как 16 пар с вопросами общего характера (например, такими, как «*Выведи все организации вместе с их названиями*»), так и 110 пар с вопросами, содержащими обращения к сущностям из графа знаний. Разработанная система одновременно ориентирована как на работу с помощью графического UI (*User Interface*), ориентированного на типовые сценарии, так и на работу с вопросами пользователя через окно текстового ввода. Поэтому в наборе данных присутствуют примеры с экземплярами сущностей и с идентификаторами IRI (*Internationalized Resource Identifier*) онтологии ИСАНД (см. таблицу). На основе упомянутых 110 пар была произведена автоматическая генерация вопросно-ответных пар путем подстановки в вопрос и SPARQL-запрос конкретных экземпляров сущностей. В результате были сформированы 1000 вопросно-ответных пар с IRI-идентификаторами из графовой базы данных и 1080 вопросно-ответных пар с именами собственными (*Name*) случайно выбранными из списка типичных имен собственных, находящихся на позициях идентификаторов.

К полученному списку 2080 пар вопрос – ответ были добавлены 16 пар с вопросами общего характера. В результате формирования набора данных получилось 2 096 вопросно-ответных пар. Из общего списка были сформированы тренировочная (1 680 пар), валидационная (208 пар) и тестовая (208 пар) выборки. Распределение вопросов по темам в валидационном и тестовом наборах такое же, как и в тренировочном.

При обучении использовался следующий шаблон промпта:

«Task: Generate SPARQL queries to query the knowledge graph based on the provided schema definition. ### Question {} ### Answer: {}».

Анализ длины полученных промптов показал, что максимальная длина для всех рассмотренных токенизаторов (см. приложение) ограничена 512 токенами. На рис. 2 представлено распределение длин промптов на основе сформированного набора данных на примере токенизатора от Qwen/Qwen2.5-7B-Instruct.

Примеры вопросно-ответных пар

Вопрос	Ответ	Вопрос	Ответ
Кто является автором публикации :publ_7808?	SELECT ?author WHERE { :publ_7808 :имеетАвтора ?author . }	Кто является автором публикации «Оптимизация динамических систем с импульсными управлениями и ударными воздействиями»?	SELECT ?author WHERE { ?publ a :Публикация; :название ?publ_name; :имеетАвтора ?author . FILTER(lcase(str(?publ_name)) = «Оптимизация динамических систем с импульсными управлениями и ударными воздействиями») }
(а) вопросно-ответная пара с IRI :publ_7808		(б) вопросно-ответная пара с подставленным конкретным экземпляром сущности «Оптимизация динамических систем с импульсными управлениями и ударными воздействиями»	

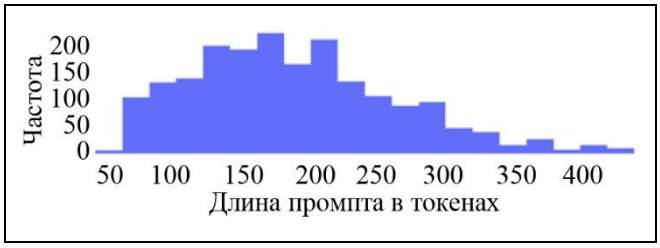

Рис. 2. Распределение длин промптов на основе сформированного набора данных для токенизатора от Qwen/Qwen2.5-7B-Instruct

3. РЕЗУЛЬТАТЫ

Для обучения LLM в данной работе использовался PEFT-подход. В рамках данного подхода были рассмотрены два метода: LoRA и QLoRA. Четырехбитная квантизация для QLoRA-адаптеров была получена с помощью Python-библиотеки bitsandbytes. Таким образом, дообучалась не LLM-модель целиком, а только часть добавленных весов, функционально выделяемых в так называемый адаптер [21].

Для решения задачи преобразования NL-to-SPARQL был рассмотрен ряд ведущих моделей, подходящих для вопросно-ответных задач и для размещения на двух видеокартах A30 из Hugging Face leaderboard [23]. Нередко академические исследования сталкиваются с проблемой нехватки вычислительных мощностей. Поэтому были рассмотрены модели со сравнительно небольшими требованиями по объему памяти графического процессора, подходящие для поставленной задачи NL-to-SPARQL-конвертера. Например, phi-3-mini-4k-i на 4 миллиарда параметров. Для еще большего снижения требований к объему видеопамяти в качестве оптимизатора в LoRA-моделях использовался paged_Adam_32bit. Для обучения четырехбитных адаптеров использовался оптимизатор Adam_torch. Другие настройки процесса обучения оставались неизменными для всех исследуемых моделей.

Для оценки результатов работы адаптеров использовалась метрика точного совпадения exact-match (EM). Из ответа LLM удалялись символы перевода строки и повторяющиеся пробелы. При exact-match-подходе сгенерированный с помощью LLM ответ должен совпадать с контрольным ответом символ в символ. Данный подход избавляет от необходимости интерпретировать сгенерированный с помощью LLM SPARQL-запрос на предмет корректного выполнения и избавляет от проведения анализа ошибок в постобработке.

На рис. 3 приведены результаты работы адаптеров LoRA и QLoRA на подготовленном авторами тестовом наборе данных. В приложении приведена таблица с результатами лучших адаптеров. Из результатов видно, что невозможно выделить одну модель, превосходящую другие по всем рассматриваемым параметрам, что говорит о наличии противоречия между требованиями качества, скорости и объема видеопамяти. В качестве практических рекомендаций можно предложить использовать наиболее быстрые модели, например, Vicuna7b-v1.5-16k (lora) или phi-3-mini-4k-i (lora) в серверных приложениях, чувствительных к скорости ответа. Модели с высоким показателем качества, например gemma-2-9b-it 4bit, разумно использовать в ситуациях, когда важна прежде всего надежность сгенерированного запроса. Наименее ресурсоемкие модели, например phi-3-mini-4k-i (4bit), потенциально применимы в мобильных приложениях.

Можно отметить общую тенденцию: четырехбитные адаптеры используют меньший объем памяти GPU, но работают медленнее. Ряд моделей позволяет на тестовом наборе данных получить качество выше 99 %, а квантизованная Google/gemma-2-9b-it успешно справилась со все-

ми вопросами из тестового набора данных. Отдельно стоит отметить, что адаптер на основе сравнительно небольшой модели phi-3-mini-4k-i на 4 миллиарда параметров хорошо справился со всеми тестовыми вопросами, использовав всего 4.4 Гб видеопамяти.

На рис. 4 представлен пользовательский интерфейс QA, который позволяет реализовать цепочку NL→SPARQL для графа знаний на основе онтологий. В верхней части интерфейса расположено поле для ввода запроса на естественном языке (например, «Кто является автором публикации Матричная теорема о лесах и лапласовские матрицы орграфов?»), для которого автоматически формируется SPARQL-запрос (кнопка Generate). Сгенерированный код отображается в редакторе SPARQL Query с нумерацией строк и подсветкой синтаксиса. Блок включает стандартные префиксы (rdfs, xsd), доменный префикс онтологии (https://www.ipu.ru/ontologies/isand#), а также тело

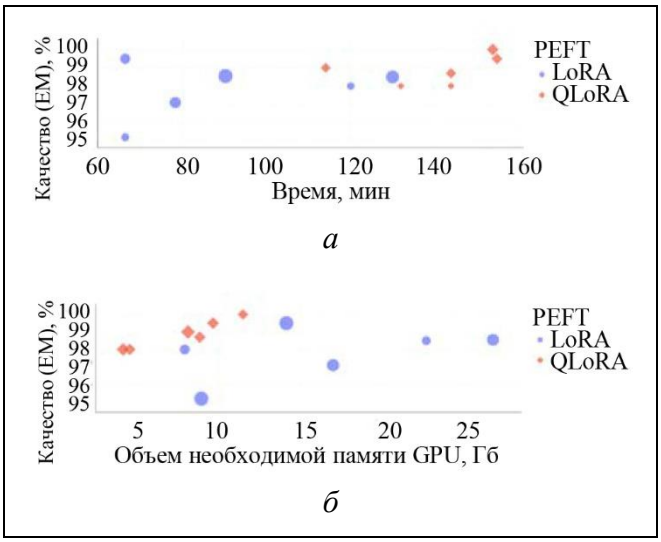

Рис. 3. Результаты работы рассмотренных моделей (каждая точка соответствует одной модели): *а* – размер фигуры отражает выделенный объем видеопамяти; *б* – размер фигуры отражает время выполнения

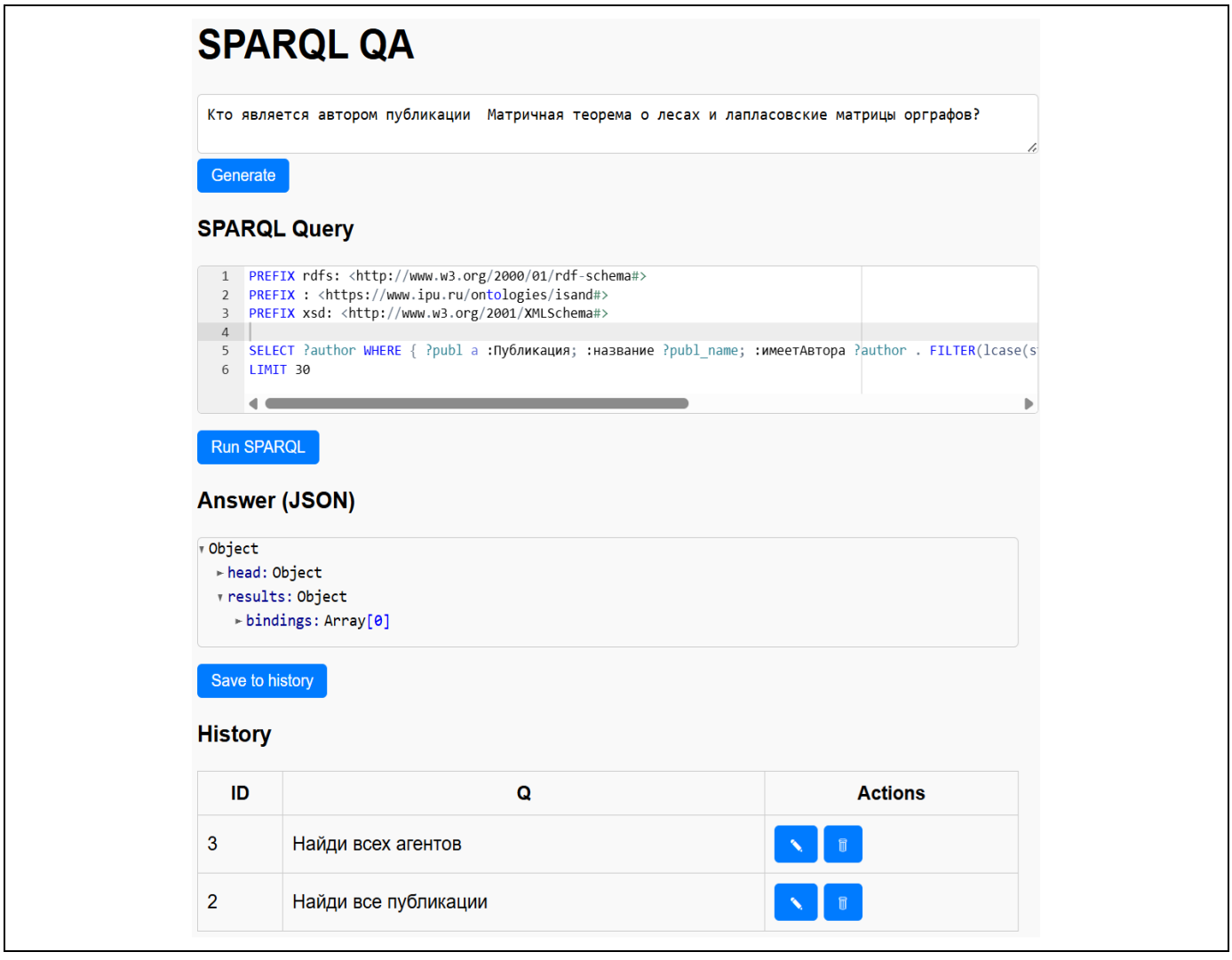

Рис. 4. Пользовательский интерфейс для выполнения запросов

запроса, построенное на классах и свойствах онтологии (например, :Публикация, :имеетАвтора, :годИздания, :фамилия, :имя). Ниже расположен блок исполнения SPARQL-запроса: кнопка «Run SPARQL» отправляет запрос на SPARQL-endpoint графовой базы данных, а панель «Answer» (JSON) возвращает машинночитаемый результат в формате JSON, что упрощает последующую обработку и верификацию. Кнопка «Save to history» сохраняет текущий вопрос (запрос) в журнал. В таблице History ведется история запросов с возможностью их редактирования и удаления. Таким образом, интерфейс одновременно поддерживает интерактивную генерацию SPARQL, прозрачное выполнение с контролируемым выводом и историю запросов.

Полученные результаты показывают, что разработанная система может успешно справляться с пользовательскими запросами, демонстрируя преимущества сочетания LLM и графовой базы знаний.

ЗАКЛЮЧЕНИЕ

Разработана гибридная вопросно-ответная система, интегрирующая возможности больших языковых моделей (LLM) и графовой базы знаний для выполнения сложных наукометрических запросов. Предложенный подход позволяет преодолеть ряд ограничений традиционных методов: вместо ручного составления формальных запросов или использования жестких шаблонов система автоматически преобразует вопрос на естественном языке в один или несколько SPARQL-запросов к базе знаний. В рамках подхода подтверждена высокая эффективность использования LLM как интерфейса к графам знаний. Показана возможность быстрой и малоресурсной адаптации LLM под конкретную онтологию.

Перспективным шагом является проверка масштабируемости решения и интеграция с более крупными и сложными графами знаний, такими как SemOpenAlex. Интерес также представляет совершенствование системы путем использования текстовых источников знаний, в первую очередь научных публикаций, для дополнения, уточнения и контекстного обоснования ответов, генерируемых системой. Наконец, по мере пополнения онтологической базы актуальными данными интерес представляет задача своевременного обновления знаний модели.

ПРИЛОЖЕНИЕ

Результаты работы адаптеров рассмотренных моделей на тестовом наборе данных

Модель	Качество (EM), %	Время выполнения (208 примеров), мин	Объем памяти GPU, необходимой для работы, Гб
Vicuna 13b-v1.5-16k lora	98,55	90	26,5
Vicuna 13b-v1.5-16k 4bit	99,5	155	9,7
Vicuna 7b-v1.5-16k lora	99,5	66	14,1
Vicuna 7b-v1.5-16k 4bit	99	114	8,2
Qwen2.5-7b-instruct lora	97,1	78	16,9
Qwen2.5-7b-instruct 4bit	98,7	144	8,9
Qwen2.5-3b-instruct lora	98	120	8
Qwen2.5-3b-instruct 4bit	98	144	4,7
Gemma-2-9b-it lora	98,5	130	22,5
Gemma-2-9b-it 4bit	100	154	11,5
Phi-3-mini-4k-i lora	95,19	66	9
Phi-3-mini-4k-i 4bit	98	132	4,3

ЛИТЕРАТУРА

1. *Wei, J., Wang, X., Schuurmans, D.,* et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models // arXiv:2201.11903. – 2022. – DOI: https://doi.org/10.48550/arXiv.2201.11903
2. *Touvron, H., Lavril, T., Izacard, G.,* et al. LLaMA: Open and Efficient Foundation Language Models // arXiv:2302.13971. – 2023. – DOI: https://doi.org/10.48550/arXiv.2302.13971
3. *Yang, A., Li, A., Yang, B.,* et al. Qwen3 Technical Report // arXiv:2505.09388. – 2025. – DOI: https://doi.org/10.48550/arXiv.2505.09388
4. *Sharma, S., Tuli, S., Badam, N.* Challenges and Applications of Large Language Models: A Comparison of GPT and DeepSeek Family of Models // arXiv:2508.21377. – 2025. – DOI: https://doi.org/10.48550/arXiv.2508.21377
5. *OpenAI* GPT-4 Technical Report // arXiv:2303.08774. – 2023. – DOI: https://doi.org/10.48550/arXiv.2303.08774
6. *Sima, A.-C., Farias, T.M.* On the Potential of Artificial Intelligence Chatbots for Data Exploration of Federated Bioinformatics Knowledge Graphs // arXiv:2304.10427. – 2023. – DOI: https://doi.org/10.48550/arXiv.2304.10427
7. *Prabhune, S., Berndt, D.J.* Deploying Large Language Models with Retrieval Augmented Generation // arXiv:2411.11895. – 2024. – DOI: https://doi.org/10.48550/arXiv.2411.11895
8. *Klager, G.G., Polleres, A.* Is GPT Fit for KGQA? // CEUR Workshop Proceedings. – 2023. – Vol. 3447. – Art. no. 11.
9. *Lopez, V., Unger, C., Cimiano, P., Motta, E.* Evaluating Question Answering over Linked Data // Web Semantics Science Services And Agents On The World Wide Web. – 2013. – Vol. 21. – P. 3–13.
10. *Färber, M.* The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data // Proceedings of 18th International Semantic Web Conference / In: Lecture Notes in Computer Science. Ed. by C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, F. Gandon. – Cham: Springer International Publishing, 2019. – Vol. 11778. – P. 113–129.
11. *Färber, M., Lamprecht, D., Krause, J.,* et al. SemOpenAlex: The Scientific Landscape in 26 Billion RDF Triples // arXiv:2308.03671. – 2023. – DOI: https://doi.org/10.48550/arXiv.2308.03671
12. *Губанов Д.А., Кузнецов О.П., Курако Е.А.* и др. Информационная система анализа научной деятельности (ИСАНД) в области теории управления // Проблемы управления. – 2024. – № 3. – С. 42–65. [*Gubanov, D.A., Kuznetsov, O.P, Kurako, E.A.,* et al. ISAND: An Information System for Scientific Activity Analysis (in the Field of Control Theory and Its Applications) // Control Sciences. – 2024. – No. 3. – P. 35–55.]
13. *Lopez, V., Fernández, M., Mooa, E., Stieler, N.* PowerAqua: Supporting Users in Querying and Exploring the Semantic Web // Semantic Web. – 2012. – Vol. 3, no. 3. – P. 249–265.
14. *Bast, H., Haussmann, E.* More Accurate Question Answering on Freebase // Proceedings of the 24th ACM International Conference on Information and Knowledge Management. – Melbourne, Australia, 2015. – P. 1431–1440.
15. *Luz, F.F., Finger, M.* Semantic Parsing Natural Language into SPARQL: Improving Target Language Representation with Neural Attention // arXiv:1803.04329. – 2018. – DOI: https://doi.org/10.48550/arXiv.1803.04329
16. *Rangel, J.C., Mendes de Farias, T., Sima, A.C. and Kobayashi, N.* SPARQL Generation: An Analysis on Fine-Tuning OpenLLaMA for Question Answering over a Life Science Knowledge Graph // arXiv:2402.04627. – 2024. – DOI: https://doi.org/10.48550/arXiv.2402.04627
17. *Fan, W., Ding, Y., Ning, L.,* et al. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models // arXiv:2405.06211. – 2024. – DOI: https://doi.org/10.48550/arXiv.2405.06211
18. *Gao, Y., Xiong, Y., Gao, X.,* et al. Retrieval-Augmented Generation for Large Language Models: A Survey // arXiv:2312.10997. – 2024. – DOI: https://doi.org/10.48550/arXiv.2312.10997
19. *Kwan, L., Omran, P. G., Taylor, K.* Using Knowledge Graphs and Agentic LLMs for Factuality Text Assessment and Improvement // CEUR Workshop Proceedings. – 2024. – Vol. 3828. – Art. no. 18.
20. *Lialin, V., Deshpande, V., Yao, X., Rumshisky, A.* Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning // arXiv:2303.15647. – 2023. – DOI: 10.48550/arXiv.2303.15647
21. *Hu, E.J., Shen, Y., Wallis, P.,* et al. LoRA: Low-Rank Adaptation of Large Language Models // arXiv:2106.09685. – 2021. – DOI: 10.48550/arXiv.2106.09685
22. *Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.* QLoRA: Efficient Finetuning of Quantized LLMs // arXiv:2305.14314. 2023. – DOI: 10.48550/arXiv.2305.14314
23. *Fourrier, C., Habib, N., Lozovskaya, A.,* et al. Open LLM Leaderboard v2. – URL: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard (дата обращения: 02.04.2025). [Accessed April 2, 2025].

Статья представлена к публикации членом редколлегии О. П. Кузнецовым.

Поступила в редакцию 23.09.2025,
после доработки 23.11.2025.
Принята к публикации 16.12.2025.

Губанов Дмитрий Алексеевич – д-р техн. наук,
✉ dmitry.a.g@gmail.com,
ORCID iD: https://orcid.org/0000-0002-0099-3386.

Сергеев Владимир Александрович – канд. техн. наук,
✉ sergeev.bureau@gmail.com,
ORCID iD: https://orcid.org/0000-0002-7948-8656

Институт проблем управления им. В. А. Трапезникова РАН, г. Москва

INTEGRATING ONTOLOGIES AND LARGE LANGUAGE MODELS FOR SCIENTOMETRIC QUESTION ANSWERING: A CASE STUDY IN CONTROL THEORY

D. A. Gubanov* and V. A. Sergeev**

*,**Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia

*✉ dmitry.a.g@gmail.com, **✉ sergeev.bureau@gmail.com

Abstract. This paper considers the problem of automatic answering complex scientometric questions formulated in natural language (NL) over knowledge bases. The study is topical due to the limitations of modern large language models (LLMs): despite high understanding capabilities, they tend to generate inaccurate responses to user questions and may have outdated information in specialized subject areas. At the same time, knowledge graphs provide accurate and relevant information but require knowledge of a formal query language. A hybrid architecture-based solution is proposed: an LLM acts as an intelligent interface to an ontology-driven knowledge base, converting NL questions into correct SPARQL queries, and the results are returned to the user. To solve the problem, a specialized data corpus is compiled to train and test NL-to-SPARQL models in the field of control theory. The approach is implemented based on the ontology of scientific activity in the field of control theory and validated on the generated corpus of questions. Integration of the LLM with the ontology-driven knowledge base ensures a high accuracy of answers (about 99%), which confirms the prospects of this approach.

Keywords: scientometrics, graph knowledge bases, large language model (LLM).