Панчехина Анна Сергеевна, студентка 4 курса бакалавриата, факультет информатики и вычислительной техники, ФГБОУ ВО «Ярославский государственный университет», РФ г. Ярославль

РАЗРАБОТКА AI-АГЕНТА ДЛЯ РАБОТЫ С БАЗАМИ ДАННЫХ

Аннотация.

В статье рассмотрены способы применения и внедрения AI-решений для работы с базами данных. Описан AI-агент, преобразующий запросы на естественном языке в SQL-код, что решает проблему нехватки компетенций и знаний у людей, не владеющих SQL. Проанализированы варианты подключения LLM, а также ключевая роль промпта. Предложенный метод позволит пользователям, эффективно получать аналитику из реляционных баз данных.

Annotation

The article discusses the ways of applying and implementing AI solutions for working with databases. It describes an AI agent that converts language queries into SQL code, which solves the problem of lack of competencies and knowledge among people who do not know SQL. The article analyzes the options for connecting LLM, as well as the key role of prompt. The proposed method will allow users to effectively obtain analytics from relational databases.

Ключевые слова: искусственный интеллект, база данных, промпт инжиниринг, большая языковая модель, запрос

Keywords: artificial intelligence, database, ptompt engineering, large language model, query

AI-агент - это система на базе генеративного искусственного интеллекта, способная планировать и совершать автономные действия во внешней среде, реагировать на изменения и взаимодействовать с человеком и другими агентами для достижения поставленных целей.

Основным инструментом для хранения огромного количества данных являются базы данных. Самым распространенным видом являются реляционные базы данных, состоящие из сущностей, атрибутов и связей. Для взаимодействия и получения информации из них необходимо знание языка SQL и написания специальных структурированных запросов. На данном этапе возникает проблема непонимания аналитиком данного инструмента и, как следствие, необходимость найма отдельного человека, либо затрачивание временных, денежных, человеческих ресурсов на обучение действующих работников. Решением данной проблемы может стать внедрение AI-агента, предназначенного для работы с базами данных.

Предлагаемая схема решения отображена на рисунке 1.

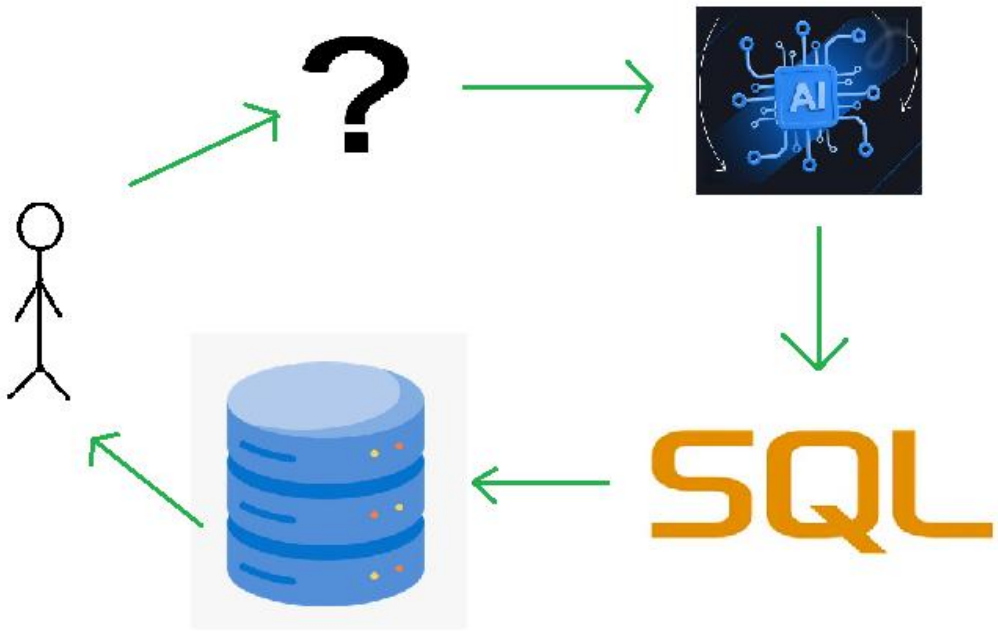

Рисунок 1 - схема работы агента.

Первым делом пользователь четко составляет свой запрос - кратко и без лишних слов формулирует ту информацию, которую он ищет. Данный запрос отправляется в нейросетевую модель, и, на основании системного и пользовательского промпта, искусственный интеллект самостоятельно

формирует SQL запрос к базе данных. Полученный запрос направляется в базу данных и выполняется. Пользователю возвращается результат SQL-запроса, то есть информация, которую он искал.

Подробнее остановимся на нейросетевых моделях.

Существует два способа подключения к нейросетевой модели. Первый из них - подключение через API. API (Application Programming Interface) предоставляет доступ к уже развернутой модели на серверах провайдера, например, Сбера, в случае с GigaChat. Данный подход не использует мощное локальное оборудование и достаточно прост в интеграции, но не может работать без стабильного интернет соединения, а также может запрашивать дополнительные платные токены на генерацию.

Второй способ - развертывание LLM-модели локально. В таком случаем модель загружается и развертывается на локальном оборудовании. Наиболее популярными являются открытые модели семейства LLaMA (Large Language Model Meta AI) и Qwen. Собственная LLM модель позволяет полностью контролировать данные, так как запросы не покидают внутренний контур. Также, преимуществом является то, что запросы можно отправлять без доступа к интернету. Благодаря локальному развертыванию, собственную модель можно постоянно предметно дообучать под конкретную специфику и терминологию для достижения более точного результата и стабильной работы. Одним из методов дообучения модели является LoRA. По этой методике не нужно переучивать модель полностью, добавляются дополнительные модули, которые и обучаются. Однако локальное развертывание требует мощного оборудования и квалифицированных специалистов для установки и поддержки инфраструктуры.

Оба способа имеют свои плюсы и минусы, выбор основывается на бюджете, времени и технических ресурсах.

Важную роль в работе агента играет промпт.

Промпт - это текстовый запрос или инструкция на естественном языке, которая задается языковой модели с целью получения нужного ответа или выполнения задачи. Промпт складывается из двух частей: системного и пользовательского сообщения. Системный промпт включает в себя общие правила поведения, задание формата ответа, рекомендации при составлении SQL-запроса. Пользовательский запрос - это сам вопрос на естественном языке, который хочет задает пользователь.

При составлении промпта важно указывать определенные элементы.

Во-первых, стоит описать структуру базы банных: перечень таблиц, атрибутов, связей. Без этого модель не сможет сгенерировать корректный SQL-запрос. Во-вторых, необходимо указать четко правила для составления запроса: какие конструкции и операторы лучше использовать. В-третьих, необходимо написать в каком формате модели стоит выдавать ответ. Это необходимо для структурированной единообразной обработки запроса в дальнейшем. Такая структура промпта позволяет добиться высокой точности преобразования естественного языка в SQL.

Подводя итог, стоит отметить, что разработанный по предложенной схеме AI-агент позволит пользователям, не владеющим SQL, получать информацию из баз данных без затруднений, задавая вопросы на естественном языке. Решение может быть внедрено под любую реляционную базу данных и использоваться для получения аналитики.

Литература

1. Грофф Джеймс Р., Вайнберг SQL: Полное руководство 2-е издание. BHW «Ирина», 2001
2. https://courses.sberuniversity.ru/llm-gigachat/1/1/1 **(**LLM: что такое большие языковые модели**)**
3. https://www.anthropic.com/engineering/building-effective-agents (Создание эффективных агентов)
4. https://www.promptingguide.ai/ru (Руководство по промпт инжинирингу)

5. https://petr-panda.ru/chto-takoe-lora-doobuchenie/ (Что такое LoRA дообучение?)

Literature

1. Groff James R., Weinberg Paul N. SQL: The Complete Reference, 2nd edition. BHW "Irina", 2001.
2. https://courses.sberuniversity.ru/llm-gigachat/1/1/1 (LLM: What are Large Language Models)
3. https://www.anthropic.com/engineering/building-effective-agents (Building effective agents)
4. https://www.promptingguide.ai/ru (Prompting Guide)
5. https://petr-panda.ru/chto-takoe-lora-doobuchenie/ (What is LoRA retraining?)