

УДК 528.3:004.896

РАЗРАБОТКА СПЕЦИАЛИЗИРОВАННОГО AI-АССИСТЕНТА ДЛЯ ИНЖЕНЕРОВ-ГЕОДЕЗИСТОВ С ИСПОЛЬЗОВАНИЕМ RAG-ПОДХОДА И ВЕКТОРНЫХ БАЗ ДАННЫХ

В.А. СКАКУН, В.В. ГОРОВЕЦ
(Представлено: М.В. Волошина)

В статье рассматривается актуальная проблема недостаточной персонализации и адаптации универсальных AI-решений, таких как ChatGPT, к профессиональным задачам. В ответ на существующие ограничения предложена концепция и реализован прототип специализированного AI-ассистента для инженеров-геодезистов.

В последние годы искусственный интеллект все активнее внедряется в здравоохранение, образование, промышленность и другие сферы. Такие инструменты, как ChatGPT, помогли компаниям оптимизировать рабочие процессы, улучшить взаимодействие с клиентами и повысить общую эффективность. Это ускорило развитие целых отраслей, включая IT, кибербезопасность и маркетинг, где AI-инструменты используются для автоматизации поддержки, создания контента и множества других задач. Однако с ростом интереса к ChatGPT компании начали сталкиваться с его ограничениями. Среди ключевых проблем – перегрузка сети в часы пиковых нагрузок, ограниченное количество бесплатных запросов и отсутствие узкоспециализированных функций.

Одна из самых важных проблем – недостаточная персонализация контента. Ответы, генерируемые ChatGPT, часто ограничиваются общими фразами и не содержат специфики, необходимой для выполнения сложных задач. Кроме того, стиль текста может создавать ощущение «роботизированной» речи. Пользователи замечают повторяющиеся синтаксические конструкции, кальки с английского языка и лексические шаблоны, которые характерны для модели, обученной преимущественно на англоязычных данных. Из этого можно сделать вывод, что для написания технической документации или обработки сложных данных специализированные инструменты окажутся более точными и эффективными [1]. Вследствие чего возникает задача создания специализированного AI-ассистента, который будет выполнять свои функции в определенной области.

AI-ассистент – это система на основе искусственного интеллекта (ИИ), которая взаимодействует с пользователем через голосовые или текстовые команды. Примером популярных виртуальных помощников являются Siri от Apple, Alexa от Amazon, Google Assistant от Google и Алиса от Yandex. Эти программы помогают пользователям управлять устройствами, искать информацию, планировать встречи, контролировать умный дом и т.д. AI-ассистентов можно разделить на несколько основных групп:

1. Чат-боты – это помощники с искусственным интеллектом, которые взаимодействуют с пользователями через интерфейсы чата.

2. Голосовые ассистенты наподобие Алисы или Ок Google отвечают на голосовой запрос.

3. Аватары с искусственным интеллектом – это графические или анимированные изображения помощников с искусственным интеллектом. Они обычно используются в приложениях и веб-сайтах для обеспечения интерактивного и визуально привлекательного взаимодействия.

4. Специализированные виртуальные помощники предназначены для оказания поддержки в конкретных отраслях или задачах. Например, виртуальный ассистент в сфере геодезии может помогать обрабатывать данные дистанционного зондирования Земли [2].

Существует множество платформ, позволяющих создавать и внедрять AI-агентов:

1) n8n – платформа no-code автоматизации, где можно собирать агентов с доступом к API, ИИ и логике принятия решений.

2) Make – визуальный конструктор сценариев. Позволяет связывать ИИ-модели, базы данных, чаты и CRM-системы в автоматические цепочки.

3) LangChain и CrewAI – для построения сложных логик на базе LLM. Подходят тем, кто работает с продвинутыми сценариями.

4) AutoGPT, Superagent, IBM Watsonx, Botpress – инструменты для создания более автономных решений с возможностью обучения и масштабирования [3].

В статье предлагается создание AI-ассистента для инженеров-геодезистов. Цель AI-ассистента: помогать инженерам-геодезистам в быстром поиске ответов на возникающие вопросы, с использованием действующих нормативных документов Республики Беларусь. Для поиска ответов AI-ассистент использует способ RAG. RAG (Retrieval Augmented Generation) (Генерация, дополненная поиском) – это способ работы с нейросетями, при котором генерация ответа LLM (большой языковой модели) осуществляется на основе данных, полученных в результате поиска в собственной базе данных (знаний). Для начала работы был создан Telegram Bot «GeoNorm» с помощью сервиса BotFather. В качестве платформы для автоматизации процессов был использован сервис n8n.

Для подключения базы знаний использовался сервис Pinecone, который является платформой векторной базы данных, разработанной для поддержки приложений искусственного интеллекта (ИИ). Все сервисы были объединены на платформе n8n. На рисунке представлена схема работы разработанного нами прототипа AI-ассистента для инженера-геодезиста (рисунок 1).

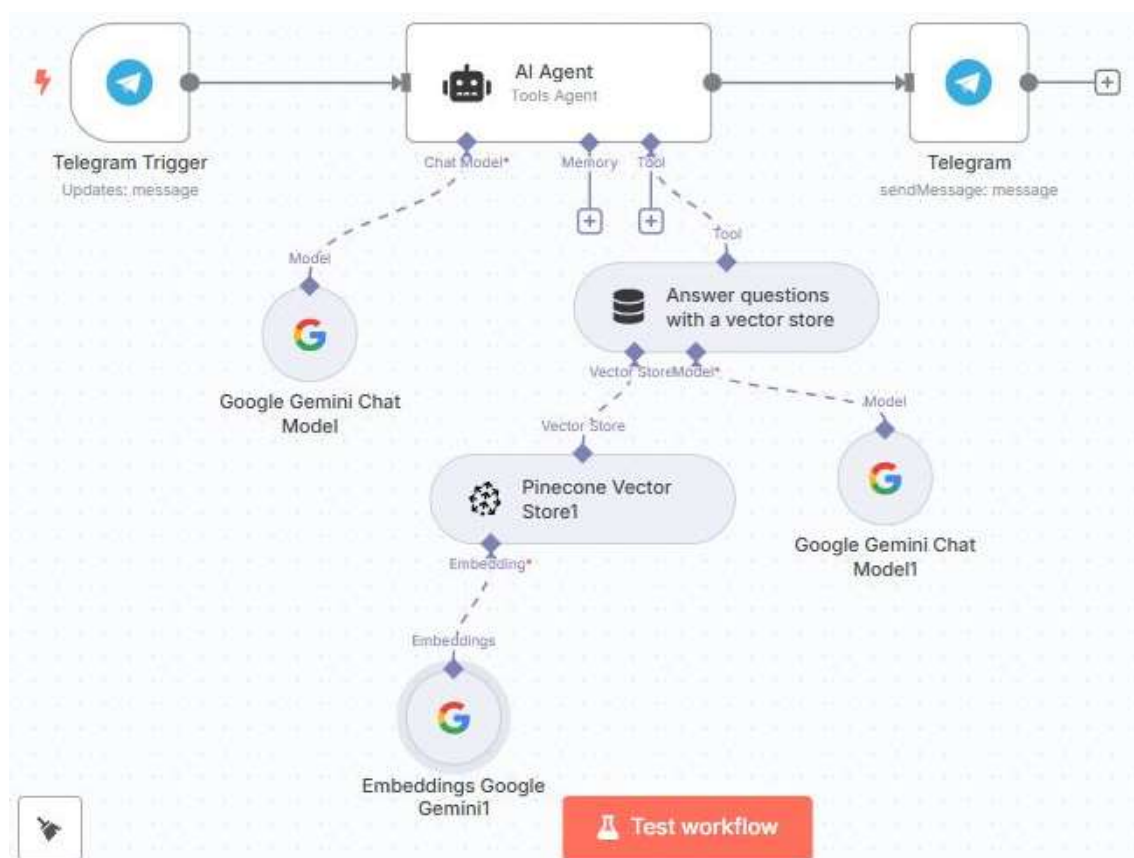


Рисунок 1. – Схема работы AI-ассистента

Схема работает по следующему принципу. В телеграмм бот поступает сообщение от пользователя с вопросом, в сервисе n8n срабатывает триггер, и текст сообщения попадает в ноду AI Agent в которой текст проверяется на соответствие требованиям (для AI Agent написан промт в котором указана роль ассистента, т.е. данный ассистент является помощником для инженера-геодезиста и отвечает на вопросы связанные с геодезией, топографией, инженерными изысканиями, расписаны правила и требования к ответам на вопросы, запрещен ответ на вопросы не связанные с геодезией, расписаны действия если вопрос на ответ не будет найден в базе данных).

Далее нода AI Agent обращается к векторной базе данных, которая содержит информацию, на основе которой AI-ассистент будет отвечать на вопросы.

Для создания базы данных (базы знаний) был создан дополнительный проект в сервисе n8n. Его суть заключается в том, чтобы любые файлы преобразовывать в векторные данные и загружать в векторную базу данных Pinecone.

Нами был загружен нормативный документ Республики Беларусь СН 1.02.01-2019 Инженерные изыскания для строительства [4] в специальную ноду, которая с помощью модели нейросети Google Gemini преобразовывает текст в векторы и загружает в виде таблицы на сервис Pinecone. Это делается по причине того, что нейросети не умеют читать наши тексты в общечеловеческом представлении. Например, отправляя запрос в любую из нейросетей, происходит следующее:

1. Нейросеть принимает текст;
2. Трансформирует этот текст в векторы (эмбединги);
3. Использует эти эмбединги для генерации ответа на основе обученных параметров модели;
4. Делает вычисления;
5. Получает ответ в виде эмбединга;
6. Трансформирует эмбединг в человеческий текст (Natural Language);
7. Отправляет нам человеческий текст.

Если коротко, то векторная база данных – это представление текстовых или байтовых данных в числовом формате, точнее – в виде векторов. Для примера, слово «привет» в векторном представлении будет иметь следующий вид $[-0.012, 0.124, -0.056, 0.203, \dots, 0.078]$ [5]. К примеру, при обычном поиске информации в тексте, будет происходить поиск точных совпадений слов, т.е. запрос «автомобиль» не найдет «машина». При поиске по эмбедингам осуществляется сопоставление слов по смыслу: «автомобиль», «машина», «транспорт» будут рядом в числах, и база их свяжет.

Загруженный PDF-файл с нормативным документом, в котором содержалось 112 страниц занял в базе данных 684 строки, в каждой из которой по 700 символов.

На следующем этапе AI Agent обращается к векторной базе данных, в которой модель нейросети Google Gemini ищет эмбединги и затем передает эти данные другой модели, которая уже оформляет ответ. После этого текст с ответом отправляется пользователю в чат Telegram Bot «GeoNorm».

Был проведен эксперимент, в рамках которого был задан один и тот же вопрос ChatGPT и созданному нами AI-ассистенту «GeoNorm», работающему на базе Telegram (рисунок 2). Результаты подтвердили нашу гипотезу: ChatGPT предоставил общий, неконкретный ответ, который не всегда соответствовал действительности, тогда как наш AI-ассистент дал точный и корректный ответ в соответствии с нормативным документом. Это стало возможным благодаря его обучению на специализированных данных по геодезии и использованию собственной базы знаний для генерации ответов.

Геодезической основой при инженерно-геодезических изысканиях в плановом положении служат (для ответа используй СН 1.02.01-2019):

Согласно СН 1.02.01-2019 «Инженерные изыскания для строительства»,

геодезической основой при инженерно-геодезических изысканиях в плановом положении служат:

- пункты государственной геодезической сети (ГГС);
- пункты местной (производственной) геодезической сети;
- пункты съемочного обоснования, создаваемого специально для изысканий на объекте.

Эти пункты обеспечивают координатное закрепление съемки и служат опорой для точного определения планового положения объектов на местности.

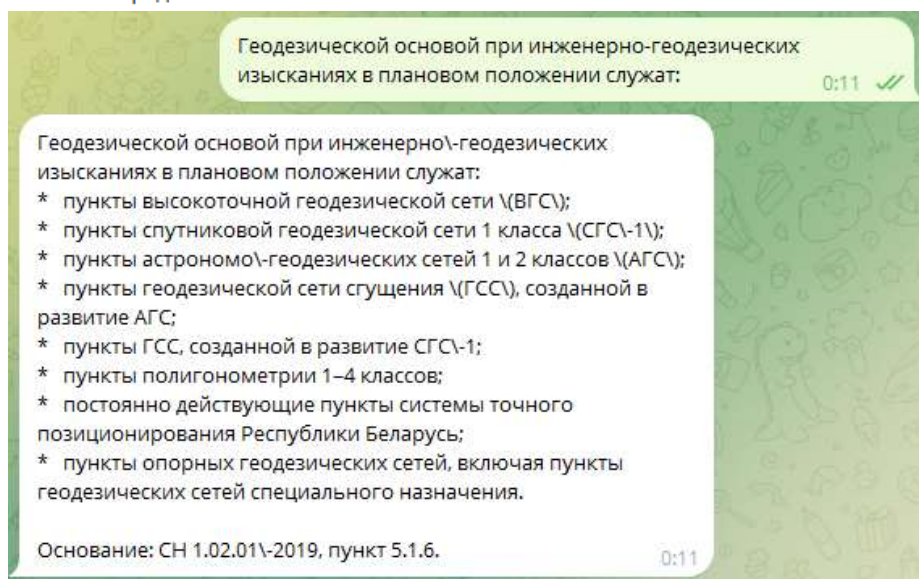


Рисунок 2. – Пример ответа ChatGPT (сверху) и ответа AI-ассистента «GeoNorm» в Telegram боте (снизу)

Преимущества использования AI-ассистентов:

1. Экономия времени и ресурсов;
2. Повышение продуктивности и уменьшение ошибок;
3. Индивидуализация и адаптация;
4. Доступность 24/7.

В результате проделанной работы был реализован прототип специализированного AI-ассистента «GeoNorm», ориентированного на нужды инженеров-геодезистов. Использование RAG-подхода и векторной базы данных позволило обеспечить релевантные и точные ответы на основе локального нормативного документа. Такая система демонстрирует высокую эффективность в профессиональной среде, снижает затраты времени на поиск информации и повышает точность принимаемых решений. Разработка подтверждает целесообразность создания узкоспециализированных AI-решений, адаптированных к конкретной отрасли, и открывает перспективы масштабирования в смежные инженерные направления.

ЛИТЕРАТУРА

1. Вокруг да около ChatGPT: AI-ассистенты, о которых вы могли не слышать [Электронный ресурс]. – URL: <https://habr.com/ru/companies/documenterra/articles/877304/> (дата обращения: 19.09.2025).
2. Как создать своего ИИ Ассистента [Электронный ресурс]. – URL: <https://vc.ru/marketing/1411401-kak-sozdat-svoego-ii-assistenta> (дата обращения: 22.09.2025).
3. AI-агенты. Где создавать и отличия от AI-ассистентов. [Электронный ресурс]. – URL: <https://vc.ru/ai/2032193-ai-agenty-i-otlichiya-ot-ai-assistentov> (дата обращения: 29.09.2025).
4. СН 1.02.01-2019 Инженерные изыскания для строительства.
5. Личный ИИ-ассистент на ваших данных. Часть 1: Векторная база ChromaDB + DeepSeek | GPT [Электронный ресурс]. – URL: <https://habr.com/ru/companies/amvera/articles/897830/> (дата обращения: 30.09.2025).