МИНОБРНАУКИ РОССИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра физики открытых систем
наименование кафедры

Внедрение и реализация локального ассистента для работы с пользовательскими данными

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТА

студента 4 курса 4041 группы

направления 09.03.02 «Информационные системы и технологии»
код и наименование направления

Института физики
наименование факультета

Седова Никиты Артемовича
фамилия, имя, отчество

Научный руководитель

д. ф-м. н., профессор		Павлов А.Н.
должность, ученая степень, уч. звание	подпись, дата	Инициалы Фамилия

Зав. кафедрой физики открытых систем
полное наименование кафедры

д.ф.-м.н., профессор		А.А.Короновский
должность, ученая степень, уч. звание	подпись, дата	Инициалы Фамилия

Саратов 2026 г.

Введение

Актуальность работы связана с быстрым развитием языковых моделей и ростом потребности в инструментах, которые помогают пользователю работать не с абстрактной информацией из сети, а с собственными файлами, заметками, документами, программным кодом и другими локальными данными. Обычный поиск по ключевым словам уже не всегда справляется с такой задачей, поскольку пользователь часто формулирует запрос естественным языком и ожидает не просто найденный фрагмент, а связный ответ по содержанию конкретных материалов.

Особое значение приобретает локальный формат работы интеллектуального ассистента. Облачные сервисы удобны, но при обработке персональных, учебных, рабочих или служебных данных возникает вопрос контроля над информацией. Передача документов во внешнюю инфраструктуру не всегда допустима, особенно если файлы содержат личные сведения, внутренние материалы организации, фрагменты кода или результаты учебной работы. Локальный ассистент позволяет снизить такую зависимость, поскольку основные процессы обработки выполняются в пределах пользовательской среды.

Цель выпускной квалификационной работы заключается в разработке и обосновании программного решения локального ассистента для работы с пользовательскими данными на основе языковой модели с учётом требований к архитектуре, безопасности, поиску информации и качеству формирования ответов.

Для достижения поставленной цели были решены следующие задачи. Первая задача связана с раскрытием архитектурных и теоретических основ локальных интеллектуальных ассистентов, включая развитие языковых моделей, принципы их функционирования и особенности локальной обработки пользовательских данных. Вторая задача направлена на проектирование и

реализацию программной системы, где рассматриваются выбор технологического стека, организация взаимодействия компонентов, механизм обработки запросов и интерфейс пользователя. Третья задача состоит в оценке эффективности разработанного решения, включая анализ точности ответов, скорости обработки запросов, устойчивости поиска и перспектив дальнейшего развития ассистента.

Объектом работы выступает процесс разработки локальных интеллектуальных программных систем, предназначенных для обработки пользовательских данных с применением языковых моделей. Предметом работы является архитектура, функциональная логика и программная реализация локального ассистента, который выполняет поиск, анализ и формирование ответов по пользовательским файлам в локальной среде.

Методологическую основу составили анализ предметной области, сравнение архитектурных подходов, систематизация требований к локальным ИИ-ассистентам, проектирование программной архитектуры, функциональный анализ разработанных модулей и практическая проверка работы системы. Отдельное внимание уделено обработке запросов, чтению документов разных форматов, поисковому механизму, взаимодействию с локальной языковой моделью и кэшированию подтверждённых результатов.

Практическая значимость работы состоит в возможности применения разработанного подхода для создания персонального локального ассистента, который помогает пользователю быстрее находить информацию в собственных документах и получать ответы на естественном языке без обязательной передачи данных во внешние сервисы. Разработанное решение может быть полезно в учебной, офисной и индивидуальной рабочей среде, где важны приватность, удобство взаимодействия и объяснимая работа с локальным набором файлов.

Основное содержание работы

1. Общая характеристика выпускной квалификационной работы

Выпускная квалификационная работа посвящена разработке локального ассистента для работы с пользовательскими данными на основе языковой модели. Выбор данной темы связан с тем, что современные пользователи всё чаще работают с большими массивами собственных цифровых материалов. Это могут быть учебные документы, заметки, отчёты, таблицы, программный код, архивные файлы и личные записи. Обычный поиск по названию файла или отдельному слову в такой ситуации уже не всегда удобен. Пользователю важно не просто найти документ, а получить понятный ответ по его содержанию.

В центре работы находится идея локальной интеллектуальной системы. Её отличие от облачного сервиса состоит в том, что обработка пользовательских данных выполняется внутри собственной рабочей среды. Такой подход особенно важен при работе с конфиденциальной информацией, поскольку документы не требуют обязательной передачи во внешнюю инфраструктуру. ВКР рассматривает локального ассистента не как простую оболочку вокруг языковой модели, а как модульное программное решение, где отдельно организованы поиск, чтение файлов, проверка релевантности, синтез ответа и хранение подтверждённых результатов.

Целью ВКР стала разработка и обоснование программного решения локального ассистента для работы с пользовательскими данными на основе языковой модели. Для достижения цели в работе были раскрыты теоретические основы локальных интеллектуальных ассистентов, рассмотрены принципы работы языковых моделей, выполнено проектирование архитектуры системы, реализованы основные программные компоненты и проведена оценка эффективности разработанного решения.

Объектом работы выступает процесс разработки локальных интеллектуальных программных систем, предназначенных для обработки пользовательских данных с применением языковых моделей. Предметом является архитектура, функциональная логика и программная реализация локального ассистента, который выполняет поиск, анализ и формирование ответов по пользовательским файлам в локальной среде.

2. Архитектурные и теоретические основы разработки локального интеллектуального ассистента

В первой части ВКР были рассмотрены теоретические основы построения интеллектуальных ассистентов на базе языковых моделей. Особое внимание уделено развитию технологий обработки естественного языка. В работе показано, что современные LLM стали результатом перехода от правиловых и статистических подходов к нейросетевым архитектурам, способным учитывать контекст и формировать связный ответ на естественном языке.

Ключевое значение имеет архитектура Transformer, поскольку именно она лежит в основе многих современных языковых моделей. За счёт механизма внимания модель способна учитывать взаимосвязи между отдельными токенами и работать не только с ближайшими словами, но и с более широким контекстом. Для локального ассистента это важно, поскольку пользовательский запрос часто связан не с одним словом, а с содержанием документа, смыслом фрагмента или связью между несколькими файлами.

Отдельный блок ВКР посвящён локальной обработке пользовательских данных. В нём раскрываются проблемы конфиденциальности и безопасности, преимущества локального развёртывания языковых моделей, а также ограничения подобных систем. Локальный формат позволяет сохранить контроль над файлами, снизить зависимость от внешних сервисов и сделать работу ассистента более прозрачной. При этом такой подход требует учёта

вычислительных ресурсов устройства, объёма оперативной памяти, скорости обработки запросов и размера контекстного окна.

Важным выводом теоретической части стало понимание того, что локальная система должна строиться не вокруг одной языковой модели, а вокруг согласованной архитектуры. Модель отвечает за языковую интерпретацию и формирование ответа, но перед этим система должна найти документы, выделить релевантные фрагменты, ограничить лишний контекст и передать в генерацию только проверенные материалы. Именно такая логика обеспечивает баланс между точностью, скоростью и безопасностью.

3. Проектирование и реализация локального ассистента для работы с пользовательскими данными

Практическая часть ВКР связана с проектированием и реализацией локального ассистента. В качестве основного языка программирования выбран Python. Такой выбор обусловлен развитой экосистемой библиотек, удобством работы с текстом и возможностью построить модульную структуру приложения. Для пользовательского интерфейса применена библиотека Tkinter, которая позволяет создать настольное приложение без использования браузера и дополнительной веб-инфраструктуры.

В качестве языковой модели используется qwen3:8b, подключённая через локальный Ollama API. Это решение стало важным элементом проекта, поскольку позволило организовать взаимодействие с LLM внутри локальной среды. Модель применяется не изолированно, а как часть общей системы обработки пользовательского запроса. ВКР подчёркивает, что Python обеспечил гибкость разработки, Tkinter позволил создать самостоятельное рабочее приложение, а qwen3:8b через Ollama стала центральным интеллектуальным компонентом системы.

Архитектура ассистента включает несколько взаимосвязанных модулей. Пользователь вводит запрос через графический интерфейс. Затем система анализирует формулировку, определяет тип действия, обращается к рабочей директории, читает документы, выполняет поиск и ранжирование фрагментов. После этого найденные материалы проходят дополнительную проверку, а языковая модель формирует итоговый ответ на основе подготовленного контекста.

В проекте реализована поддержка разных типов файлов. Ассистент может работать с txt, md, json, csv, log, py, docx и pdf. Это расширяет практическую ценность программы, поскольку пользователь получает возможность обрабатывать не только простые текстовые заметки, но и документы, программные файлы, таблицы и материалы в распространённых форматах. Важную роль играет локальный механизм индексации и ранжирования, который сначала отбирает наиболее релевантные фрагменты, а уже затем подключает языковую модель для интерпретации.

Отдельное значение имеет механизм semantic expansion. Пользователь не всегда формулирует запрос теми же словами, которые содержатся в документе. Поэтому система может расширять исходную формулировку, добавлять близкие по смыслу термины и повышать устойчивость поиска. Такой подход делает ассистента более гибким, особенно при работе с естественными, неполными или неточными запросами.

Интерфейс программы построен так, чтобы пользователь мог взаимодействовать с ассистентом без технических команд. Он вводит вопрос, получает ответ и при необходимости подтверждает его полезность. В системе предусмотрено кэширование подтверждённых результатов. Если файлы не изменялись, повторный ответ может быть получен быстрее. Если документ был обновлён, ранее сохранённый результат перестаёт считаться актуальным.

4. Оценка эффективности разработанного решения и перспективы его развития

В третьей части ВКР проведена оценка эффективности локального ассистента. Проверка была направлена на анализ точности ответов, скорости обработки запросов, устойчивости поиска и общей пригодности системы для работы с пользовательскими данными. Для тестирования использовался подготовленный набор из двадцати четырёх документов, включавший текстовые файлы и один документ формата DOCX. Материалы имитировали реальные пользовательские данные, включая календарные события, рабочие задачи, учебные дедлайны, финансовые записи, архивные документы и служебные материалы.

Автоматизированная оценка выполнялась с помощью отдельного программного модуля evaluate_agent.py. Внутри проверки было задано двенадцать тестовых сценариев. Для каждого сценария определялись пользовательский запрос, ожидаемый файл-источник и контрольные ключевые слова. После обработки система фиксировала итоговый ответ, время выполнения и рассчитывала точность результата.

В режиме работы с подключённой языковой моделью средняя точность составила 90,0 %, а среднее время обработки составило 7,155 секунды. Эти показатели показывают, что система способна стабильно обрабатывать большую часть пользовательских запросов и находить нужный источник данных. Наиболее высокие результаты были получены при работе с учебными дедлайнами, рабочими задачами, финансовыми документами и DOCX-файлом.

Дополнительно был проведён контрольный цикл без подключения языковой модели. В этом режиме средняя точность составила 95,0 %, а среднее время обработки составило 0,012 секунды. Сравнение двух режимов позволило выявить важную закономерность. Базовый поиск обеспечивает максимальную скорость и высокую буквальную точность, а подключение LLM делает ответы

более структурированными, логически завершёнными и удобными для восприятия.

Практическая оценка показала, что разработанный ассистент устойчиво работает при большинстве типовых пользовательских сценариев. Система корректно обрабатывает точные запросы, поддерживает работу с разными форматами документов и способна справляться с отдельными неточностями формулировок. Наибольшая чувствительность проявляется при коротких неоднозначных запросах и наличии пересекающихся архивных данных.

Перспективы развития разработанного решения связаны с расширением перечня поддерживаемых форматов, оптимизацией времени интеллектуальной генерации, улучшением работы с неоднозначными запросами и развитием персонального пользовательского контекста. Также значимым направлением может стать более гибкая настройка модели под разные сценарии применения, включая учебную, офисную и исследовательскую работу.

В итоге основное содержание ВКР показывает, что разработанный локальный ассистент является работоспособной интеллектуальной системой для обработки пользовательских данных. Он объединяет локальный поиск, чтение документов, языковую модель, пользовательский интерфейс и механизм сохранения подтверждённых результатов. Полученное решение демонстрирует практическую применимость и может рассматриваться как основа для дальнейшего развития персональных инструментов работы с локальными данными.

Заключение

В выпускной квалификационной работе была раскрыта тема разработки локального ассистента для работы с пользовательскими данными на основе языковой модели. В центре внимания находилась не просто возможность применения LLM для генерации ответов, а построение полноценного

программного решения, которое объединяет локальный поиск, обработку документов, проверку релевантности найденных фрагментов, синтез ответа и удобный пользовательский интерфейс.

Локальная архитектура решает эту задачу более безопасно, так как документы, запросы и результаты обработки остаются внутри пользовательской среды и не требуют обязательной передачи во внешние облачные сервисы.

Теоретическая часть позволила определить технологическую основу разрабатываемой системы. Были рассмотрены принципы работы современных языковых моделей, особенности архитектуры Transformer, значение контекстного окна, генеративной природы ответа и вычислительных ограничений локального запуска. Отдельное внимание было уделено вопросам конфиденциальности, поскольку именно контроль над пользовательскими файлами является одним из главных преимуществ локального ассистента.

Практическая часть работы была связана с проектированием и реализацией программной архитектуры. В качестве основного языка разработки использован Python, для интерфейса применена библиотека Tkinter, а взаимодействие с языковой моделью организовано через локальный Ollama API с моделью qwen3:8b. Пользовательский запрос поступает через интерфейс, затем проходит маршрутизацию, поиск по локальным документам, ранжирование, проверку кандидатов и только после этого передаётся в языковую модель для формирования ответа.

Значимым результатом стала поддержка работы с различными типами файлов, включая txt, md, json, csv, log, py, docx и pdf. Дополнительную практическую ценность имеет механизм semantic expansion, который повышает устойчивость поиска при несовпадении пользовательской формулировки с текстом документа.

Оценка эффективности показала, что разработанное решение демонстрирует высокий уровень практической работоспособности. В режиме с подключённой языковой моделью средняя точность составила 90,0 %, а среднее время обработки составило 7,155 секунды. Без подключения LLM точность достигла 95,0 %, а среднее время обработки составило 0,012 секунды. Базовый локальный поиск обеспечивает максимальную скорость, а языковая модель повышает качество представления результата, делает ответ более связным и понятным для пользователя.

Проведённая работа подтвердила достижение поставленной цели. Было разработано и обосновано программное решение локального ассистента для работы с пользовательскими данными на основе языковой модели. Система показала способность находить релевантные материалы в локальной директории, обрабатывать запросы пользователя, формировать ответы по найденному контексту и сохранять подтверждённые результаты.

Практическая значимость разработанного решения состоит в том, что предложенная архитектура может использоваться как основа для персональных локальных ассистентов в учебной, офисной, исследовательской и индивидуальной рабочей среде.

Перспективы дальнейшего развития связаны с расширением поддерживаемых форматов файлов, оптимизацией скорости генерации, улучшением обработки неоднозначных запросов, развитием персонального контекста и более гибкой настройкой языковой модели под разные пользовательские сценарии.

Полученный результат показывает, что локальные ассистенты на базе LLM могут стать практичным направлением развития персональных интеллектуальных систем, ориентированных на безопасную, удобную и осмысленную работу с пользовательскими данными.