МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

ПРОЕКТИРОВАНИЕ ТЕКСТОВОЙ МОДЕЛИ С КОНТЕКСТНОЙ АДАПТАЦИЕЙ ДЛЯ ЦИФРОВОГО ПОМОЩНИКА «STILLARA»

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Тищенко Максима Алексеевича

Научный руководитель:
ст. преп. кафедры ИиП _____ Казачкова А. А.

Зав. кафедрой:
к.ф.-м.н., доцент _____ Огнева М. В.

Саратов 2026

ВВЕДЕНИЕ

Развитие современных информационных технологий и автоматизация повседневных процессов приводят к изменению форматов межличностной коммуникации, повышая требования к доступности систем удаленного взаимодействия. Ограниченность каналов связи и высокая интенсивность информационных потоков обусловливают необходимость создания программных решений, способных обеспечить непрерывный текстовый диалог и симулировать естественное речевое взаимодействие.

В то же время наблюдается качественный рост эффективности алгоритмов искусственного интеллекта, в частности, больших языковых моделей (Large Language Models, LLM). Данные архитектуры обладают высокой точностью в задачах обработки естественного языка (NLP) и семантического анализа, что расширяет возможности проектирования человеко-машинных интерфейсов [1]. В отличие от диалоговых систем на основе жестких правил и конечно-автоматных сценариев, современные LLM способны поддерживать контекстную зависимость на протяжении длинных сессий, формируя основу для разработки интеллектуальных агентов с высокой степенью автономности.

Основная научно-техническая проблема при проектировании подобных диалоговых агентов заключается в высокой формализованности и шаблонности ответов базовых моделей. Оптимизация стандартных алгоритмов под фактологическую точность приводит к формированию обезличенных реплик, снижающих вовлеченность пользователя [2]. В связи с этим актуален переход от сугубо информационных критериев генерации к управлению модальностью ответов [3]. Для повышения комфорта интерфейса необходимо обеспечить адаптивность модели, ее способность к формированию недирективных поддерживающих реплик и сохранение связности контекста.

В рамках данной дипломной работы рассматривается разработка и прототипирование приложения-собеседника «Stillara», основанного на текстовой LLM-модели. Система ставит перед собой цель обеспечивать пользователям возможность непринужденного общения, выступая в роли

эмпатичного и отзывчивого виртуального компаньона. Приложение призвано создать безопасную цифровую среду для ведения диалога, способствующего рефлексии и позитивному взаимодействию.

Актуальность работы обусловлена растущим спросом на новые формы цифрового общения, необходимостью гуманизации человеко-машинных интерфейсов и потенциалом больших языковых моделей для создания по-настоящему увлекательных диалогов. Исследование и реализация возможностей LLM в области эмоционального интеллекта представляют значительный интерес для развития социально-ориентированных приложений.

Практическая значимость работы заключается в возможности интеграции разработанной эмпатической модели в состав системы цифрового помощника «Stillara», разрабатываемой в рамках междисциплинарного проекта. Вклад автора в рамках данного проекта сосредоточен на проектировании интеллектуального ядра системы — текстовой модели, обеспечивающей недирективную поддержку пользователей.

Вышесказанное позволяет сформулировать **цель работы** — спроектировать и реализовать текстовую модель на основе архитектуры LLM для её последующей интеграции в программный комплекс цифрового помощника «Stillara», способную распознавать эмоциональное состояние пользователя по текстовому запросу и генерировать эмпатический, поддерживающий ответ, строго соответствующий принципам недирективного подхода и валидации чувств пользователя.

Поставленная цель определила следующие **задачи**:

1. Проанализировать теоретические основы недирективной терапии и методы их алгоритмизации в современных LLM.
2. Изучить архитектуры больших языковых моделей и методы их параметрически-эффективного дообучения (PEFT), в частности LoRA/QLoRA.
3. Исследовать влияние аппаратных ограничений на процесс обучения и подобрать оптимальную вычислительную конфигурацию для работы с моделями различного масштаба.

4. Разработать методику формирования специализированного синтетического набора данных «Stillara Gold» с использованием цепочек рассуждений (Chain-of-Thought) и дистилляции знаний.

5. Реализовать процесс дообучения модели на базе открытой архитектуры (Qwen 2.5 7B) с внедрением скрытого поля рассуждений (thought).

6. Разработать систему комплексной оценки качества эмпатии, включающую математические метрики и качественную экспертизу по шкале Труакса-Кархаффа методом «Model-as-a-Judge».

7. Провести сравнительный анализ эффективности дообученной модели в сопоставлении с базовыми архитектурами и проприетарными решениями (GPT-4o).

Методологические основы проектирования и адаптации диалоговых систем с контекстной зависимостью представлены в работах ведущих отечественных и зарубежных исследователей в области искусственного интеллекта, обработки естественного языка и гуманистической психологии. Теоретические основы недирективного взаимодействия, феноменологии общения и терапевтической валидации чувств, легшие в основу разработки метрик оценки качества эмпатии (включая шкалу Труакса-Кархаффа), базируются на фундаментальных трудах К. Роджерса (C. Rogers). Технические и архитектурные аспекты разработки диалоговых трансформерных систем опираются на исследования таких ученых, как Васвани А. (A. Vaswani) (архитектура Transformer), Радфорд А. (A. Radford) (авторегрессионные GPT-системы), Ху Э. (E. Hu) (методология низкоранговой адаптации LoRA), Деттмерс Т. (T. Dettmers) (технологии 4-битного квантования весов QLoRA), Вей Дж. (J. Wei) (концепция цепочек рассуждений Chain-of-Thought), а также Шнейдерман Б. (ориентированный на человека безопасный ИИ).

Практическая значимость бакалаврской работы заключается в возможности интеграции разработанной эмпатической модели в состав системы цифрового помощника «Stillara», разрабатываемой в рамках междисциплинарного проекта. Вклад автора сосредоточен на проектировании

интеллектуального ядра системы — текстовой модели, обеспечивающей предсказуемую недирективную поддержку пользователей с соблюдением стилистических и этических ограничений без передачи персональных данных сторонним операторам.

Структура и объём работы. Бакалаврская работа состоит из введения, 6 разделов, заключения, списка использованных источников и 7 приложений. Общий объем работы – 83 страницы, из них 67 страниц – основное содержание, включая 2 рисунка и 10 таблиц, цифровой носитель в качестве приложения, список использованных источников информации – 29 наименований.

КРАТКОЕСОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретические основы проектирования и адаптации диалоговых систем с контекстной зависимостью» посвящен анализу решений в области разработки диалоговых языковых моделей с эмоциональным интеллектом. Выявлена проблема проприетарных систем (Inflection Pi, Google Gemini, OpenAI GPT-4o) — «барьер совета» (алгоритмы RLHF заставляют модель навязывать директивные инструкции вместо текстового присутствия). Обоснован выбор мультиязычного семейства Qwen 2.5 как наиболее лингвистически устойчивого для русскоязычной среды. Описан математический аппарат эффективной параметрической настройки PEFT (LoRA, QLoRA), минимизирующий эффект катастрофического забывания базовой сети, а также изучены способы оценки качества эмпатии методом Model-as-a-Judge по шкале Труакса-Кархаффа.

Второй раздел «Методологические основы проектирования текстового модуля и формирование первичных данных» посвящен определению базовых методик оценки параметров терапевтического диалога. Описаны операционные критерии недирективного подхода К. Роджерса и принципы валидации чувств, накладывающие запрет на императивные суждения. Сформулированы требования к текстовому корпусу, способному переобучить авторегрессионный вывод с прямого распределения вероятностей на двухэтапный инференс с явным формированием скрытого аналитического поля

рассуждений (thought).

Третий раздел «Техническая реализация и масштабирование системы» отражает процесс оптимизации моделей под аппаратные ограничения. Описан локальный этап разработки на базе GPU с 8 ГБ VRAM (DirectML, Windows). Зафиксированы критические отказы (OOM) полноразмерных моделей и исследован инференс малых архитектур в соответствии с таблицей 1.

Таблица 1 — Результаты предварительного тестирования базовых моделей на локальном аппаратном стеке

Архитектура модели	Количество параметров	Ожидаемый VRAM (float16)	Результат тестирования	Причина отказа / Статус
Mistral-7B	7.2 млрд	~14.5 ГБ	Критический отказ	Превышение объема ОЗУ и видеопамяти
Phi-3-mini	3.8 млрд	~7.6 ГБ	Ошибка аллокации	Нехватка памяти для буферов активации
Qwen2-1.5B	1.5 млрд	~3.1 ГБ	Успешный запуск	Оптимальный баланс ресурсов

Четвертый раздел «Разработка архитектурных решений, алгоритмов дообучения и структуры представления данных диалоговой системы» посвящен программной реализации цикла Fine-tuning модели Qwen 2.5 7B в среде Unsloth. Внедрена спецификация формата разметки ChatML, использующая системные токены для жесткой изоляции логических ролей («system», «user», «thought», «assistant»). Описана конфигурация SFT-трейнера с применением 8-битного оптимизатора AdamW, накоплением градиентов (gradient_accumulation_steps = 4) и косинусным планировщиком скорости обучения.

Пятый раздел «Разработка высококачественного набора данных методом синтетической дистилляции» описывает конвейер создания специализированного датасета «Stillara Gold». Метод прямой генерации был отвергнут из-за риска «галлюцинаций эмпатии». Реализован двухступенчатый

процесс: синтез через ролевое моделирование и Chain-of-Thought на базе GPT-4o (с обязательной генерацией внутреннего монолога психолога) и кросс-валидация независимой моделью-судьей (Model-as-a-Judge). Из корпуса безвозвратно удалялись реплики с оценкой ниже 4 баллов по шкале Труакса-Кархаффа. Параметры репрезентативности и объемы полученных данных по эмоциональным категориям приведены в таблице 2.

Таблица 2 — Количественное распределение и репрезентативность эмоциональных категорий в наборе данных «Stillara Gold»

Эмоция	Количество примеров	Статус выборки
Беспокойство	166	Репрезентативно
Уныние	155	Репрезентативно
Грусть	148	Репрезентативно
Радость	88	Оптимально
Возбуждение	68	Достаточно
Счастье	64	Достаточно
Удовлетворение	60	Достаточно
Злость	53	Достаточно
Непонимание	25	Требует внимания

Шестой раздел «Анализ результатов и оценка эффективности дообучения» содержит количественные и качественные результаты проведенных вычислительных экспериментов. Математически доказана высокая эффективность выбранного метода контролируемой оптимизации весовых коэффициентов на разработанном целевом корпусе. Динамика изменения математических показателей модели на различных стадиях обучения зафиксирована в таблице 3.

Таблица 3 — Динамика показателей обучения

Стадия обучения	Значение Loss	Перплексия (PPL)
Начальное состояние (Base)	3.16	23.57
Финальное состояние (Stillara)	0.81	2.25

Важнейшим качественным результатом работы стало успешное формирование скрытого аналитического поля рассуждений (thought) в процессе двухэтапного инференса трансформера. Наличие данного поля позволяет

модели перед генерацией итогового ответа провести латентный семантический и психологический анализ реплики пользователя. В таблице 4 продемонстрирован практический пример реализации внутреннего монолога модели при обработке скрытой фрустрации пользователя.

Таблица 4 — Пример реализации внутреннего монолога при обработке скрытой фрустрации

Поле	Содержимое (Логика модели)
Thought	Эмоция: сарказм, скрытая злость/разочарование. Анализ: пользователь использует иронию, чтобы скрыть глубокую фрустрацию от череды неудач. Прямое сочувствие («Мне жаль») может быть воспринято как издевка. Стратегия: признать тяжесть дня, не «подыгрывая» сарказму, а давая пространство для выхода реальных эмоций.
Output	«Когда всё рушится одновременно, ирония иногда остается единственным способом защиты. Похоже, сегодня на тебя свалилось слишком много, и это действительно выматывает. Что стало той самой "последней каплей" в этом дне?»

ЗАКЛЮЧЕНИЕ

В ходе выполнения бакалаврской работы была успешно достигнута поставленная цель по проектированию и реализации текстовой модели для цифрового помощника «Stillara». Итогом стал программный прототип на базе архитектуры Qwen 2.5 7B, способный проводить глубокий контекстный анализ пользовательских запросов и генерировать недирективные ответы, направленные на корректную интерпретацию интенций пользователя и удержание конструктивного диалога.

В теоретической части работы были детально проанализированы принципы построения недирективных диалоговых стратегий и методы их алгоритмизации в современных языковых моделях. Исследование

существующих архитектур и методов параметрически-эффективного дообучения показало, что технология низкоранговой адаптации (LoRA) в сочетании с 4-битным квантованием формата NF4 является наиболее оптимальным способом адаптации весовых коэффициентов под узкоспециализированные лингвистические задачи в условиях фиксированных вычислительных ресурсов.

Особое внимание было уделено исследованию влияния аппаратных спецификаций на процесс моделирования. В результате серии вычислительных экспериментов был обоснован и сформирован эффективный программно-аппаратный стек. На этапе предварительного анализа были определены лимиты локальных графических процессоров на базе архитектуры RDNA с использованием библиотеки DirectML, что обусловило последующий переход к облачной инфраструктуре и применение оптимизированных алгоритмов обучения с помощью библиотек Unsloth и PEFT. Данный комплекс решений позволил преодолеть ограничения видеопамяти и успешно провести Fine-tuning модели класса 7B, обеспечив высокую стабильность градиентного спуска.

Ключевым этапом практической реализации стала разработка методики формирования специализированного синтетического набора данных «Stillara Gold». Применение метода цепочек рассуждений (Chain-of-Thought) позволило зафиксировать процесс многокритериального семантического анализа внутри скрытых слоев модели. Процесс дистилляции знаний из проприетарной архитектуры высокого порядка GPT-4o обеспечил высокую вариативность и грамматическую корректность обучающей выборки, а внедрение обязательного поля thought формализовало алгоритм построения логического вывода нейросети перед генерацией итоговой реплики.

Разработанная система комплексной оценки, интегрирующая математические метрики и автоматизированную экспертизу методом «Model-as-a-Judge», подтвердила валидность предложенного подхода. Проектирование экспертного фильтра на базе модифицированной шкалы

Труакса-Кархаффа позволило объективизировать процесс кросс-валидации данных и минимизировать риск генерации шаблонных или директивных конструкций. Сравнительный анализ показал, что дообученная модель «Stillara» существенно превосходит базовые решения в точности следования заданным коммуникативным нормам и качестве контекстной адаптации, демонстрируя показатели, сопоставимые с крупными закрытыми системами общего назначения.

Важно отметить, что разработка данной системы не ограничилась рамками академического исследования, а стала технологическим ядром для реального стартап-проекта. Практические результаты работы были успешно представлены и защищены в рамках дополнительной образовательной программы «Стартап как диплом». В ходе защиты экспертной комиссией была подтверждена не только техническая состоятельность и глубина программного решения, но и его высокая потенциальная востребованность на рынке интеллектуальных цифровых ассистентов с поддержкой недирективного взаимодействия.

Таким образом, все задачи, поставленные во введении, были решены в полном объеме. Работа имеет выраженную практическую значимость для проектирования интеллектуальных диалоговых стратегий через скрытые цепочки рассуждений и открывает новые перспективы для повышения автономности и качества работы цифровых помощников в рамках проекта «Stillara».

Отдельные результаты бакалаврской работы и архитектурные решения цифрового помощника **были успешно представлены**, защищены и высоко оценены в рамках междисциплинарной образовательной программы «Стартап как диплом» (Саратовский национальный исследовательский государственный университет имени Н.Г. Чернышевского, 2026 г.).

Основные источники информации:

1. Роджерс К. Р. Клиентоцентрированная психотерапия: ее современная практика, контекст и теория. – М.: Рефл-бук, 2020. – 320 с.

2. Васавани А. и др. Внимание - это все, что вам нужно // Труды конференции NIPS. – 2017. – Т. 30. – С. 84-95. [84-90]
3. Ху Э. Дж. и др. LoRA: низкоранговая адаптация больших языковых моделей // Материалы международной конференции ICLR. – 2022. – С. 1-25. [1-15]
4. Деттмерс Т. и др. QLoRA: эффективное тонкое обучение квантованных LLM // Труды конференции NeurIPS. – 2023. – Т. 36. – С. 30-45. [30-40]
5. Wei J. et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [Электронный ресурс] // Advances in Neural Information Processing Systems. – 2022. – Т. 35. – С. 24824-24837. – URL: https://arxiv.org/abs/2201.11903
6. Расчкин Х. и др. К эмпатическим моделям открытых диалогов: новый бенчмарк и датасет // Труды 57-й ежегодной встречи Ассоциации компьютерной лингвистики. – 2019. – С. 210-225. [210-220]
7. Lee, Y. et al. Evaluation of Generative Language Models for Empathic Conversational Agents // arXiv preprint arXiv:2401.03456. – 2024. – URL: https://arxiv.org/abs/2401.03456
8. Шнейдерман Б. Искусственный интеллект, ориентированный на человека: надежность, безопасность и доверие. – М.: ДМК Пресс, 2023. – 420 с. [50-80, 120-150]
9. Чжан Й. и др. DialogPT: крупномасштабное генеративное предварительное обучение для генерации диалоговых ответов // Труды 58-й ежегодной встречи Ассоциации компьютерной лингвистики. – 2020. – С. 150-165. [150-160]