КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2026 Т. 18 № 1 С. 41–59
DOI: 10.20537/2076-7633-2026-18-1-41-59

Ки&М

МОДЕЛИ В ФИЗИКЕ И ТЕХНОЛОГИИ

УДК: 004.85

Ресурсно-адаптивный подход к разметке текстовых данных в структурированном виде с использованием малых языковых моделей

С. А. Антипова[1,a], А. М. Журкин[2,b]

[1]Военная академия материально-технического обеспечения имени генерала армии А. В. Хрулёва,
Россия, 199034, г. Санкт-Петербург, наб. Макарова, д. 8
[2]Санкт-Петербургский политехнический университет Петра Великого,
Россия, 195251, г. Санкт-Петербург, ул. Политехническая, д. 29

E-mail: [a] samiraspb11@gmail.com, [b] zhurkin_al@list.ru

Получено 04.11.2025, после доработки — 24.12.2025.
Принято к публикации 04.02.2026.

В данной работе проведено экспериментальное исследование применения автоматической разметки текстовых данных в формате «вопрос – ответ» (QA-пары) в условиях ограниченных вычислительных ресурсов и требований к защите данных. В отличие от традиционных подходов, основанных на жестких правилах или использовании внешних API, предложено применять малые языковые модели с небольшим количеством параметров, способные функционировать локально без GPU на стандартных CPU-системах. Для тестирования были выбраны две модели: Gemma-3-4b и Qwen-2.5-3b (квантованные 4-битные версии), а в качестве исходного материала использован корпус документов с четкой структурой и формально-строгим стилем изложения. Разработана система автоматической аннотации, реализующая полный цикл генерации QA-датасета: автоматическое разбиение исходного документа на логически связные фрагменты, формирование пар «вопрос – ответ» моделью Gemma-3-4b, предварительная проверка их корректности с использованием Qwen-2.5-3b с опорой на доказательный фрагмент из контекста и экспертной оценкой качества. Экспорт полученных результатов предоставляется в формате JSONL. Оценка производительности охватывает всю систему генерации QA-пар, включая обработку фрагментов локальной языковой моделью, модули предобработки и постобработки текста. Производительность измеряется по времени генерации одной QA-пары, общей пропускной способности системы, использованию оперативной памяти и загрузке процессора, что позволяет объективно оценить вычислительную эффективность предлагаемого подхода при запуске на CPU. Эксперимент на расширенной выборке из 12 документов показал, что автоматическая аннотация демонстрирует устойчивую производительность при обработке документов различных типов, тогда как ручная разметка характеризуется существенно большими временными затратами и высокой вариативностью. В зависимости от типа документа ускорение аннотации по сравнению с ручным процессом составляет от 8 до 14 раз. Анализ качества показал, что большинство сгенерированных QA-пар обладают высокой семантической согласованностью с исходным контекстом, при этом лишь ограниченная доля данных требует экспертной корректировки или исключения. Хотя полная ручная валидация корпуса (золотой стандарт) в рамках работы не проводилась, сочетание автоматической оценки и выборочной экспертной проверки позволяет рассматривать полученный уровень качества как приемлемый для задач предварительной автоматизированной аннотации. В целом результаты подтверждают практическую применимость малых языковых моделей для построения автономных и воспроизводимых систем автоматической разметки текстов в условиях ограниченных вычислительных ресурсов и создают основу для дальнейших исследований в области эффективной подготовки обучающих корпусов для задач обработки естественного языка.

Ключевые слова: языковые модели, разметка данных, вопрос – ответ, оценка качества, локальные вычисления, ограниченные вычислительные ресурсы

COMPUTER RESEARCH AND MODELING
2026 VOL. 18 NO. 1 P. 41–59
DOI: 10.20537/2076-7633-2026-18-1-41-59

Ки&М

MODELS IN PHYSICS AND TECHNOLOGY

UDC: 004.85

Resource-adaptive approach to structured text data annotation using small language models

S. A. Antipova[1,a], A. M. Zhurkin[2,b]

[1]Military academy of logistics named after general of the army A. V. Khrulyov,
8 Makarova embankment, St. Petersburg, 199034, Russia
[2]Peter the Great St. Petersburg Polytechnic University,
29 Politekhnicheskaya st., St. Petersburg, 195251, Russia

E-mail: [a] samiraspb11@gmail.com, [b] zhurkin_al@list.ru

Received 04.11.2025, after completion — 24.12.2025.
Accepted for publication 04.02.2026.

This paper presents an experimental study of the application of automatic annotation of text data in the question – answer format (QA pairs) under conditions of limited computing resources and data protection requirements. Unlike traditional approaches based on rigid rules or the use of external APIs, we propose using small language models with a small number of parameters that can function locally without a GPU on standard CPU systems. Two models were selected for testing — Gemma-3-4b and Qwen-2.5-3b (quantized 4-bit versions) — and a corpus of documents with a clear structure and a formally rigorous style of presentation was used as source material. An automatic annotation system was developed that implements the full cycle of QA dataset generation: automatic division of the source document into logically connected fragments, formation of "question – answer" pairs using the Gemma-3-4b model, preliminary verification of their correctness using Qwen-2.5-3b based on evidence span from the context and expert quality assessment. The results are exported in JSONL format. Performance evaluation covers the entire QA pair generation system, including fragment processing by the local language model, text preprocessing and postprocessing modules. Performance is measured by the time it takes to generate a single QA pair, the total throughput of the system, RAM usage, and CPU load, which allows for an objective assessment of the computational efficiency of the proposed approach when running on a CPU. An experiment on an extended sample of 12 documents showed that automatic annotation demonstrates stable performance when processing different types of documents, while manual annotation is characterized by significantly higher time costs and high variability. Depending on the type of document, the acceleration of annotation compared to the manual process ranges from 8 to 14 times. Quality analysis showed that most of the generated QA pairs have high semantic consistency with the original context, with only a limited proportion of data requiring expert correction or exception. Although full manual validation of the corpus (the "gold standard") was not performed as part of this work, the combination of automatic evaluation and selective expert review allows us to consider the resulting quality level acceptable for preliminary automated annotation tasks. Overall, the results confirm the practical applicability of small language models for building autonomous and reproducible automatic text annotation systems under limited computational resources and provide a basis for further research in the field of effective training corpus preparation for natural language processing tasks.

Keywords: language models, data annotation, question – answer, quality evaluation, local computation, limited computational resource

Citation: *Computer Research and Modeling*, 2026, vol. 18, no. 1, pp. 41–59 (Russian).

1. Введение

Современные интеллектуальные системы — от поисковых агентов и рекомендательных платформ до автономных роботов и аналитических модулей военного назначения — в значительной степени зависят от наличия качественно размеченных данных. Разметка (аннотация) является неотъемлемым этапом построения обучающих выборок, на которых обучаются модели обработки естественного языка, компьютерного зрения, мультимодальные трансформеры. Стремительная эволюция больших языковых моделей (Large Language Models, далее — LLMs) и систем генерации данных с расширенным поиском (Retrieval-Augmented Generation, далее — RAG) также неразрывно связана с прогрессом в методологии подготовки данных. Однако традиционная ручная разметка, будучи золотым стандартом качества, становится узким местом из-за своей дороговизны и медлительности, а в некоторых случаях — из-за необходимости привлечения экспертов в специфических предметных областях. Также следует отметить, что в условиях ограниченных вычислительных мощностей и невозможности использовать внешние API к публичным LLM (ChatGPT, Claude, DeepSeek, Gemini и т. п.) возникает задача автономной и ускоренной разметки текстов без доступа к облачным сервисам и GPU. Традиционные подходы на основе извлечения именованных сущностей (Named Entity Recognition, NER) и лингвистических правил требуют ручной адаптации под предметную область и снижают скорость аннотации. В ответ на этот вызов научное сообщество активно исследует и внедряет новые, более автоматизированные и эффективные подходы.

В работе [Schroeder et al., 2025] авторы провели качественный обзор и систематизацию исследований, которые используют LLM именно в задачах аннотирования и синтеза данных. Использовать гибридные методы — сочетать человеческую аннотацию и LLM, особенно на начальных стадиях, для контроля качества. При синтезе данных важно разнообразие: не просто генерировать однотипные инструкции/примеры, но и обеспечивать вариативность запросов, а также всегда предусматривать этап модерации и отбора результатов разметки — не полагаться на все, что сгенерировано моделью.

Тем не менее в большинстве существующих исследований, включая работы 2023–2025 гг. [Zhuang et al., 2023; Shi et al., 2024; Jahan et al., 2024; Xia et al., 2025], генерация аннотаций осуществляется с опорой на внешние API проприетарных LLM (например, ChatGPT, Claude, Gemini, Mistral-Large), обладающих высокой вычислительной мощностью и обширными контекстными окнами. Подобные подходы, хотя и обеспечивают высокое качество синтетических данных, невоспроизводимы в условиях ограниченных вычислительных ресурсов, а также неприемлемы в защищенных доменах, где использование облачных сервисов противоречит требованиям конфиденциальности (военные, медицинские, юридические, промышленные системы и др.).

В работе [Busta, Oyler, 2025] демонстрируется, что малые языковые модели (до 8B параметров, Small Language Models, далее — SLM) способны эффективно выполнять задачи извлечения и структурирования информации из научных текстов без использования крупных LLM или внешних API. Авторы применяют компактные модели для автоматического выделения сущностей и атрибутов (по сути — аннотаций) в корпусе биологических публикаций. Показано, что при корректной постановке задачи и инструкциях SLM достигают высокой точности и стабильности, существенно снижая вычислительные и финансовые затраты по сравнению с крупными моделями. Этот принцип позволяет снизить необходимость в ручной разметке данных при решении задач, требующих глубоких знаний, и ускорить адаптацию LLM к новым доменам и информации, повышая ее способности к интеграции знаний и рассуждению. Подобный подход демонстрирует, что SLM можно использовать для самостоятельного создания обучающих данных, которые одновременно отражают фактические знания и логические шаблоны рассуждений, открывая новые возможности для развития более продвинутых рассуждающих способностей в мощных системах искусственного интеллекта.

Тем не менее данное направление — одно из самых слабо проработанных областей современной практики аннотирования данных. Работы, рассматривающие автономную аннотацию с локальными моделями, немногочисленны. Генерация QA — это более сложная задача, чем простая классификация или продолжение текста. Модель должна сначала проанализировать и понять текст, затем выделить ключевую информацию, сформулировать осмысленный вопрос к ней и только потом сгенерировать точный ответ. Выполнить всю эту цепочку автономно (или с минимальным участием человека) и качественно на слабом «железе» — открытая проблема.

В условиях невозможности использования мощных моделей и внешних API остаются пока недостаточно исследованными следующие фундаментальные научные проблемы и связанные с ними инженерные задачи.

Проблема качества автономной генерации: «легковесные» модели (особенно квантованные) не обладают достаточной «рассуждающей» способностью, чтобы полностью *автономно* (без экспертного надзора) генерировать сложные, разнообразные и фактически корректные QA-пары для специализированных научно-технических текстов с высокой степенью смысловой плотности, поскольку такие модели склонны к упрощению логики, пропуску контекстных связей и генерации поверхностных или фактически неточных ответов, требующих экспертной валидации и дообработки. Малые модели (до 7B) часто не обладают достаточной когнитивной глубиной для стабильного и разнообразного построения вопросов и ответов, особенно без внешней поддержки (API, сильных публичных моделей). Как следствие, имеет место повышенная доля логических ошибок и банальных вопросов, а также тематически размытых и неточных ответов [Kim et al., 2025]. Тем не менее применимость различных малых языковых моделей к задаче корректной аннотации документов с определенной строгой структурой (нормативно-правовые акты, распоряжения, ГОСТы, справочные документы и т. д.) возможна, но пока недостаточно экспериментально раскрыта в современных научно-прикладных исследованиях, особенно в русскоязычных изданиях.

Сложность применения безреференсных метрик оценки качества QA-пар (Perplexity, LLM-as-a-judge, NLI-Score и др.) для сложных нормативно-правовых (с узкоспециализированной направленностью) документов. Обзорные статьи по метрикам оценки генерации текста [Gehrmann et al., 2023; Gao et al., 2025; Sindhujan et al., 2025] неизменно приходят к выводу, что безреференсные метрики пока не могут надежно заменить человеческую оценку или эталоны, так как оценивают лишь *отдельные аспекты* (например, семантическую близость или согласованность), но не качество в целом (например, релевантность или нетривиальность вопроса). Так или иначе, необходимы комбинация различных метрик для повышения надежности оценки в разнообразных задачах обработки естественного языка и адаптация кастомного, интегрального решения под свои данные.

Недостаток исследований по оптимизации процесса аннотации для слабого железа. Оптимизация процесса аннотации на слабом железе включает не только саму модель, но и весь пайплайн: разбиение на фрагменты, эффективность операций ввода-вывода при чтении этих фрагментов, выбор размера батча (пакета) для подачи в модель и алгоритмы постобработки полученных аннотаций, направленные на минимизацию потребления RAM и нагрузки на CPU, а также формулировку запроса.

К вышесказанному следует добавить следующее. Во-первых, эффективно разбивать большие документы для моделей с малым контекстным окном (4–8K токенов) — отдельная инженерная задача, требующая баланса между размером фрагментов, их смысловой связностью и перекрытием контекста. Необходимо обеспечить, чтобы каждый фрагмент содержал достаточно информации для корректного ответа модели, но при этом не выходил за пределы контекстного окна. Для этого применяются довольно известные методы semantic chunking (семантическое разбиение по смысловым блокам (чанкам)), context overlap (перекрытие соседних фрагментов

на 10–20 %), а также динамическое разбиение на основе структуры документа (заголовков, абзацев, маркеров списка). В системах RAG подобное разбиение критично: оно напрямую влияет на точность поиска, релевантность ответов и эффективность использования вычислительных ресурсов.

А, во-вторых, «легковесные» модели крайне чувствительны к формулировкам текстовых запросов (в том числе системных промтов). Подбор промтов, которые стабильно работают для QA-генерации именно на моделях с $\leqslant$ 7B параметров, требует отдельных исследований, так как такие модели часто демонстрируют нестабильность на уровне понимания и обобщения контекста, склонность к «галлюцинациям» и сильную зависимость от структуры вопроса. Эффективные промты для них должны быть максимально конкретными, структурированными и контекстно насыщенными, но при этом не перегруженными избыточной информацией [Errica et al., 2025; Son, Kim, 2025].

Также следует отметить, что большинство практических руководств по развертыванию больших языковых моделей на ресурсно-ограниченных вычислительных платформах в основном ориентированы на задачи инференса и оптимизации вывода (компрессия, ускорение), тогда как вопросы построения полноценных пайплайнов генерации и аннотирования данных остаются слабо формализованными, что подтверждается анализом профильных журнальных публикаций и обзорных работ [Zhen et al., 2025; Zheng et al., 2025].

Целью данного исследования является экспериментальная проверка подхода к автоматической аннотации данных в формате «вопрос – ответ» без использования публичных LLM, обеспечивающего приемлемое качество QA-пар при ограниченных вычислительных ресурсах.

Для достижения цели исследования решаются следующие задачи:

1) определение требований к системе автоматической аннотации QA-пар, работающей на ограниченных вычислительных ресурсах без обращения к публичным LLM;

2) реализация прототипа программного комплекса, обеспечивающего генерацию QA-пар с использованием локальных малых языковых моделей;

3) проведение серии практических экспериментов по оценке скорости и качества аннотации документов в структурированном виде;

4) сравнение результатов работы системы с ручной разметкой и определение уровня приемлемости автоматической аннотации.

Научная новизна исследования заключается в следующем. Впервые экспериментально проверяется гипотеза о том, что малые языковые модели размером 3–4 млрд параметров в квантованных форматах (Q4/Q5) способны обеспечивать устойчивую генерацию QA-пар приемлемого качества при полностью локальном запуске на CPU без использования GPU и внешних API. В отличие от существующих работ, опирающихся на облачные LLM или модели $\geqslant$ 7B параметров, в данной работе исследуется полный автономный пайплайн аннотации, включающий сегментацию текста, генерацию QA-пар, автоматическую верификацию и экспертную модерацию. К известным ранее результатам относятся применение LLM для синтеза QA-датасетов и использование QA-основанных метрик фактической согласованности. Собственный вклад работы состоит в адаптации этих подходов к условиям ограниченных вычислительных ресурсов, экспериментальной оценке качества и производительности локальных SLM на русскоязычных нормативных текстах, а также в разработке ресурсно-адаптивного программного комплекса для практической аннотации документов.

Следовательно, эксперимент должен показать, насколько SLM способны обеспечивать устойчивую и воспроизводимую генерацию QA-пар, и определить оптимальный компромисс между качеством, скоростью и аппаратными затратами.

2. Экспериментальная подготовка

Для проведения экспериментов с автоматической разметкой текстовых данных и локальным применением языковых моделей использовалась персональная вычислительная система, обеспечивающая выполнение задач генерации и верификации данных без применения графических ускорителей (GPU).

Основу аппаратной конфигурации составляет процессор Intel Core i9-12900PF (архитектура *Alder Lake*, техпроцесс 10 нм), включающий 16 физических ядер (8 высокопроизводительных P-ядер и 8 энергоэффективных E-ядер) и поддерживающий до 24 потоков. Тактовая частота P-ядер достигает 4,9 ГГц, что обеспечивает высокую скорость выполнения матричных операций, характерных для инференса языковых моделей.

Система оснащена 64 ГБ оперативной памяти DDR5, работающей с эффективной частотой ≈ 4800 МГц, что обеспечивает достаточную пропускную способность для размещения и обработки весов моделей размером до 13 млрд параметров в квантованных форматах.

Экосистема llama.cpp/GGUF на сегодняшний день является де-факто стандартом локального инференса языковых моделей: квантования Q4/Q5 обеспечивают разумный баланс качества/памяти, позволяя запускать 2–7B модели и на 8–16 ГБ RAM.

Полный пайплайн разработанной системы аннотации включает следующие этапы: подготовка данных → автосегментация на фрагменты → генерация QA по каждому фрагменту → → самопроверка генерирующей моделью или проверка сторонней моделью → экспертная оценка «бракованных пар» → JSONL-экспорт — единый путь, согласованный с локальными квантованными стеками (llama.cpp/GGUF).

Для обеспечения обоснованного выбора малой языковой модели был проведен **предварительный** этап отбора, направленный на выявление моделей, демонстрирующих наилучшие показатели качества и устойчивости при решении задачи автоматизированной разметки текстов в формате «*вопрос – ответ – фрагмент контекста* (*evidence_span*)». Результаты данного этапа использовались для обоснованного сужения круга рассматриваемых моделей и последующего детального анализа только наиболее перспективных решений.

Для сравнения были выбраны открытые модели с числом параметров до 4 млрд: Qwen2.5-3B-Instruct, Gemma-3-4B-IT, Llama-3.2-3B-Instruct, Ministral-3B-Instruct и Phi-4-Mini-Instruct (табл. 1). Указанные модели доступны в виде квантованных сборок (Q4/Q5), что обеспечивает их использование на рабочих станциях и ноутбуках с ограниченными вычислительными ресурсами. Одним из ориентиров отбора также были значения бенчмарков MMLU-Pro (0-shot, CoT; техника, при которой модели предоставляется только описание задачи без каких-либо примеров, и она полагается исключительно на свои предобученные знания для генерации ответа, но при этом с пошаговым рассуждением) и его мультиязычной версии Multilingual-MMLU (5-Shot; с 5 примерами для демонстрации) для оценки способности моделей к пониманию и решению задач на различных языках, взятые из технических отчетов разработчиков моделей.

Для оценки применимости моделей к решению задачи разметки русскоязычных текстовых документов в формате «вопрос – ответ – фрагмент контекста (evidence_span)» был проведен сравнительный эксперимент на небольшом эталонном датасете. Отбор проводился на фиксированном наборе из 200 текстовых чанков, сформированных из нормативно-методических документов одного жанра и из одной предметной области с четкой структурой (обязательно в тексте присутствуют заголовки, абзацы, нумерация и таблицы) и формально-строгим стилем изложения, которые используются в сферах, где критически важны точность, однозначность, значимость и воспроизводимость. Для каждого чанка в эталонной разметке были заданы две независимые QA-пары, что в совокупности составило 400 эталонных записей.

Таблица 1. Основные характеристики отобранных для эксперимента моделей

№	Модель	Параметры, млрд	MMLU-Pro (0-shot, CoT), %	Multilingual-MMLU (5-Shot), %	Максимальный размер контекстного окна
1	Phi-4-Mini	3,8	**52,8**	49,3	128K
2	Qwen2.5-3B	3,0	44,7	**55,9**	32K
3	Gemma-3-4B	4,0	43,6	н/д*	128K
4	Llama-3.2-3B	3,0	39,2	48,1	128K
5	Ministral-3B	3,0	35,3	46,4	256K

* В техническом отчете Gemma 3 (Google) значение метрики Multilingual MMLU-Pro не приведено — в разделе multilingual указываются другие мультиязычные бенчмарки. Ближайший аналог по смыслу из официального набора — Global-MMLU-Lite со значением 57,0.

Все сравниваемые модели обрабатывали идентичный набор чанков при фиксированном запросе и неизменных параметрах генерации, которые приведены далее по тексту. Каждая модель генерировала по две QA-пары для каждого чанка без дополнительного дообучения. Главным критерием отбора являлась способность модели корректно выполнить аннотирование с первой попытки, без необходимости дополнительных уточняющих или корректирующих промтов.

Поскольку модель генерирует несколько QA-пар без фиксированного порядка, перед оценкой качества определялось наиболее вероятное соответствие между сгенерированными и эталонными парами, чтобы сравнение производилось по эквивалентным смысловым элементам.

Оценка качества разметки проводилась по двум ключевым компонентам. Корректность ответа оценивалась с использованием метрики F1 (ответа), вычисляемой между сгенерированным и эталонным ответами после нормализации текста. Итоговые значения определялись как макро-средние по всем 400 QA-парам. Дополнительно анализировались 10-й перцентиль распределения значений F1 и доля ответов, превышающих пороговое значение 0,80, что позволяло оценить устойчивость моделей на сложных примерах (табл. 2). Также следует отметить, что качество формулировки вопроса не оценивалось отдельной метрикой, поскольку для одного и того же чанка допустимы разные эквивалентные по смыслу варианты вопроса, а лексические метрики (EM (Exact Match)/F1) могут давать заниженную оценку. В рамках настоящего эксперимента задача отбора моделей фокусировалась на проверяемых и однозначно сопоставимых компонентах разметки — корректности ответа и точности evidence span, которые напрямую измеряются F1-метриками по эталону.

Таблица 2. Результаты оценки качества разметки на примере выбранных моделей

№	Модель	F1 (ответа) (сред±откл)	F1 (ответа) P10/P50/P90	F1 (фрагмента) (сред ± откл)	F1 (фрагмента) P10/P50/P90
1	**Qwen2.5-3B**	**0,84 ± 0,09**	**0,72/0,85/0,94**	**0,81 ± 0,10**	**0,69/0,82/0,93**
2	**Gemma-3-4B**	**0,86 ± 0,08**	**0,74/0,87/0,95**	**0,83 ± 0,09**	**0,71/0,84/0,94**
3	Phi-4-Mini	0,78 ± 0,14	0,55/0,81/0,92	0,76 ± 0,13	0,58/0,78/0,90
4	Llama-3.2-3B	0,71 ± 0,18	0,42/0,74/0,90	0,69 ± 0,17	0,45/0,71/0,88
5	Ministral-3B	0,69 ± 0,19	0,40/0,72/0,89	0,66 ± 0,18	0,43/0,68/0,87

Качество доказательного фрагмента оценивалось с использованием метрики F1 (фрагмента), вычисляемой по пересечению токенов между эталонным и сгенерированным evidence span. Данная метрика отражает как полноту, так и избыточность выделяемого фрагмента. Агрегация значений выполнялась аналогично метрике ответа; пороговое значение составляло 0,75.

По результатам эксперимента было установлено, что модели Qwen2.5-3B-Instruct и Gemma-3-4B-IT демонстрируют наиболее высокое и устойчивое качество разметки по обеим

F1-метрикам. Для указанных моделей наблюдались более высокие значения средних F1, а также более высокие значения 10-го перцентиля, что свидетельствует о стабильной работе на текстовых фрагментах.

Остальные рассмотренные модели характеризовались либо снижением средней корректности ответа, либо нестабильностью при выборе доказательного фрагмента, что проявлялось в пониженных значениях F1 и падении 10-го перцентиля. Кроме того, моделью Phi-4-Mini-Instruct на русскоязычном материале зафиксированы односложные ответы и грамматические ошибки в формулировке вопросов, что снижает ее удобство для прямого использования в автоматизированной разметке без дополнительного постконтроля. Модели Llama-3.2-3B-Instruct и Ministral-3B-Instruct продемонстрировали заметно более низкое качество аннотирования: в их выводах часто встречались англоязычные термины и «галлюцинации». В ряде случаев такие ответы оказались бы непригодны для прямого включения в обучающие выборки без ручной доработки.

Для проверки статистической значимости различий между моделями применялся непараметрический критерий Уилкоксона для связанных выборок, поскольку все модели оценивались на одном и том же наборе из 400 QA-пар. Различия по метрикам F1 (ответа) и F1 (фрагмента) между отобранными моделями и остальными участниками эксперимента оказались статистически значимыми ($p < 0{,}05$).

Таким образом, по результатам проведенного эксперимента модели Qwen2.5-3B-Instruct (Alibaba) и Gemma-3-4B-IT (Google) были отобраны для дальнейшего использования в задаче автоматизированной разметки текстовых документов в формате «вопрос – ответ – фрагмент контекста (evidence_span)».

3. Результаты и обсуждение

В рамках исследования разработан прототип программного комплекса, предназначенного для автоматизированного формирования пар «вопрос – ответ» (QA-пар) на основе исходных текстовых данных. Основной целью создания данного решения являлось обеспечение возможности локальной генерации, проверки и модерации QA-пар в условиях ограниченных вычислительных ресурсов и требований к защите информации.

Система реализует предобработку документов, включающую этап разбиения исходного текста на смысловые фрагменты (чанки) фиксированного или динамического размера. Это позволяет обеспечить корректную сегментацию длинных документов, формирование устойчивого контекстного окна и повышение релевантности сгенерированных QA-пар. Механизм предобработки предусматривает автоматическое определение границ смысловых блоков, фильтрацию малозначимых сегментов и привязку каждого чанка к исходному документу для последующей трассировки данных.

Программный комплекс оснащен графическим интерфейсом пользователя (GUI) (рис. 1), реализованным на базе библиотеки PySide6 (Qt for Python). Интерфейс обеспечивает удобную визуальную работу с текстовыми данными и результатами генерации, включая:

- загрузку и предварительный просмотр документов,
- настройку параметров предобработки (размер чанков, перекрытие контекста),
- запуск генерации QA-пар и отслеживание хода выполнения,
- интерактивную модерацию и корректировку результатов,
- экспорт итоговых данных в форматах JSONL.

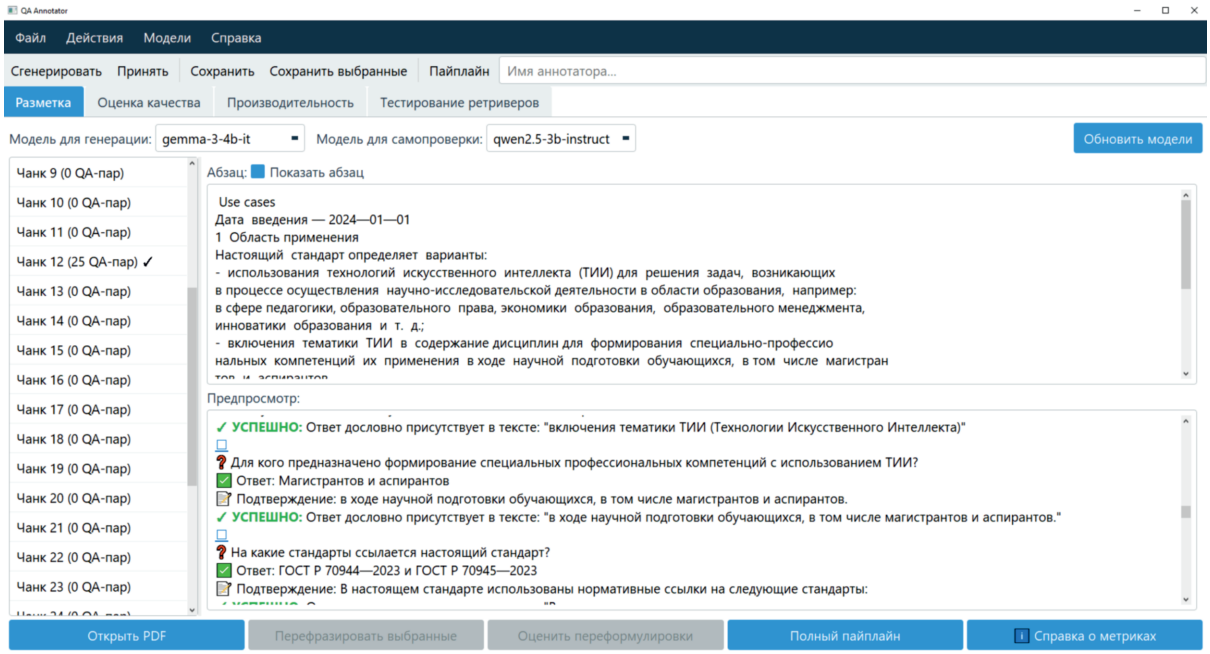

Рис. 1. Фрагмент графического пользовательского интерфейса разработанного программного комплекса

Архитектура программного комплекса спроектирована как модульная и кастомизируемая. Пользователь может самостоятельно изменять параметры генерации (температуру, длину контекста, число пар и т. д.), выбирать используемую модель инференса, адаптировать шаблоны вопросов и подключать дополнительные инструменты верификации в виде иных языковых моделей. Такая архитектура делает прототип гибким инструментом, легко адаптируемым под специфические задачи аннотирования, обучения языковых моделей и анализа текстовых данных в различных предметных областях.

Интеграция предобработки, локального инференса и визуального интерфейса в единую систему обеспечивает баланс между производительностью, автономностью и удобством работы, что делает данный программный комплекс практическим решением для задач автоматической генерации и оценки QA-пар.

3.1. Оценка производительности

Время ручной разметки зависит от нескольких факторов, но есть эмпирические данные из работ по корпусной лингвистике, подготовке датасетов QA и аннотированию нормативных документов. Например, средняя производительность варьируется от 30 до 60 QA-пар в час при поверхностной аннотации (например, выделение сущностей или вопросно-ответных пар общего характера) до 15–35 QA-пар в час при глубокой смысловой разметке нормативных или технических документов, где требуется точная интерпретация контекстных зависимостей, терминов и логико-семантических связей [Ore et al., 2021].

Для экспериментальной выборки были отобраны документы различной жанрово-структурной принадлежности при соблюдении следующих критериев:

- официальный или научно-технический стиль,
- объем — 10–25 страниц,
- наличие явной внутренней структуры (заголовки, абзацы, нумерация),
- отсутствие графических элементов, не обрабатываемых моделью.

Экспериментальный корпус включал 12 документов трех типов (нормативные (ГОСТ), инструкции, научные статьи). Объем документов варьировался от 10 до 25 страниц, а общее число сгенерированных QA-пар — от 50 до 95 на документ. Все измерения выполнялись при фиксированной аппаратной конфигурации и неизменных параметрах пайплайна, что обеспечивало сопоставимость результатов.

При запуске генерации QA-пар моделью Gemma-3-4b в эксперименте фиксировались: время до первого токена (TTFT, с), средняя скорость генерации (токенов/с), среднее время на один выходной токен (TPOT, с/токен), общее время аннотации документа и число сгенерированных QA-пар. Основные настройки модели и системный промт приведены в таблице 3.

Таблица 3. Основные параметры генерации для модели Gemma-3-4b

Параметры	Значения для QA-генерации	Системный промт
Максимальная длина ответа (в токенах)	2048	Создай формальные, нейтральные вопросы и точные ответы к ним исключительно на русском языке от специалиста широкого профиля по заданному фрагменту текста. Для этого: 1) внимательно прочитай текст на русском языке; 2) найди в тексте один важный факт или правило, который можно спросить в виде вопроса; 3) сформулируй формальный, нейтральный вопрос: полное предложение, без разговорных слов, без эмоциональной окраски; 4) вопрос должен быть сформулирован так, чтобы на него можно было ответить одним четким фрагментом из текста; 5) не добавляй информацию, которой нет в тексте (запрещены догадки и внешние знания); 6) ответ должен быть кратким, точным, строго на основе текста, без добавления несуществующих, выходящих за рамки контекста данных; 7) не используй другие языки, кроме русского; 8) не копируй текст целиком — допускается частичное цитирование, но в ответе не должно быть длинного неосмысленного копирования; 9) из текста вопроса должно быть понятно, о чем идет речь; 10) формат вывода строго jsonl: {"question": "...", "answer": "...", "evidence_span": "...", "type": "basic"}; 11) в поле вывода "evidence_span" должен содержаться конкретный отрывок из контекста, который служит обоснованием ответа
Температура сэмплирования (temperature)	0,3	
Вероятностное усечение (top_p)	0,8	
Отбор *k* наиболее вероятных токенов (top_k)	40	
Штраф за повторение	1,1	
Количество лучей при поиске (num_beams)	3	
Режим досрочного завершения генерации (early_stopping)	True	

Эксперимент повторялся трижды при фиксированном seed; в таблице 4 приводятся обобщенные показатели производительности автоматической и ручной аннотации (по категориям документов). В таблице значения показателей представлены в виде среднего значения по документам соответствующего типа со стандартным отклонением, характеризующем разброс. Документ рассматривался как независимая единица анализа.

Абсолютное время автоматической аннотации одного документа находилось в диапазоне 8,6–18,4 мин, тогда как ручная аннотация аналогичных документов требовала от 85 до 190 мин. Таким образом, во всех рассмотренных случаях автоматическая разметка обеспечивала существенное сокращение времени обработки по сравнению с ручной разметкой.

Для сопоставимости результатов дополнительно была выполнена нормализация времени аннотации в пересчете на одну QA-пару. Анализ показал, что среднее время генерации одной

Таблица 4. Показатели производительности автоматической и ручной аннотации

Категория документа	Количество документов	QA-пар (всего)	Время аннотации моделью, мин (сред ± откл)	Время ручной аннотации, мин (сред ± откл)	Диапазон
Нормативные документы	4	355	13,6 ± 1,8	150,0 ± 32,6	×9–13
Инструкции и регламенты	4	420	14,2 ± 3,9	128,5 ± 36,0	×8–11
Научные статьи	4	255	14,1 ± 3,3	148,5 ± 21,2	×11–14
Итого/среднее	**12**	**1030**	**14,0 ± 3,0**	**142,3 ± 31,0**	**×9–13**

QA-пары составляет порядка 8–10 с для нормативных документов и инструкций и увеличивается до 12–15 с для научных статей. Разброс времени в пересчете на одну QA-пару обусловлен вариативностью длины генерируемых ответов и доказательных фрагментов, а также наличием постоянной составляющей (инициализация, пакетная обработка, операции ввода-вывода), вклад которой заметнее на документах с меньшим числом QA-пар.

Так, в качестве примера, на разметку 16-страничного документа (≈ 3300 слов) ГОСТ Р 70949-2023 «Применение искусственного интеллекта в научно-исследовательской деятельности» из эксперимента у аннотатора в ручном режиме ушло 135 мин (2 ч 15 мин) и получено 75 пар, т. е. в среднем затрачено 1,8 мин на 1 QA-пару. При использовании Gemma-3-4B среднее время генерации 5 QA-пар составило 52 с, т. е. 10,4 с на одну пару, что эквивалентно 13,0 мин для документа из 75 пар. При этом TTFT = 7,4 с (на пару), TPOT = 0,025 с/токен, средняя скорость — 40 токенов/с. Таким образом, после выдачи первого токена на генерацию оставшейся части ответа тратится ≈ 3,0 с, что соответствует ∼ 120 выходным токенам на пару (при указанной скорости).

Несмотря на то что установка значения num_beams > 1 влияет на время генерации (при значении 3 время генерации примерно увеличивается на 15 %), настройка альтернативных гипотез при поиске позволяет отображать наиболее «человеческие» формулировки QA-пар. Так или иначе, предложенный подход обеспечил ускорение процесса аннотации более чем в 10 раз на примере документа серии ГОСТ (рис. 2).

Скорость аннотации в первую очередь определяется длиной и сложностью текстовых фрагментов. При переходе от нормативных документов к научным статьям наблюдается умеренное увеличение времени генерации, однако асимптотические характеристики инференса и порядок величины производительности остаются неизменными. На рис. 3 в качестве примера приведено сравнение времени ручной и автоматизированной аннотации 20-страничной научной статьи в области информационных технологий. Общее время генерации 95 QA-пар составило 18,4 мин при автоматической аннотации и 168 мин при ручной разметке, что соответствует ускорению процесса аннотации в 9,1 раза.

Эксперимент также показал, что при генерации QA-пар было задействовано 12 потоков центрального процессора, средняя загрузка CPU составляла около 65–75 %, при этом использование оперативной памяти не превышало 6 ГБ. Полученные значения показывают, что даже при работе исключительно на CPU модели размером 3–4 млрд параметров обеспечивают приемлемую производительность аннотации — порядка 40 токенов/с. Это позволяет использовать предложенный подход в локальных условиях без привлечения GPU-ресурсов.

3.2. Оценка качества Q-пар

При отсутствии эталонных ответов (золотого стандарта) классические метрики вроде Precision/Recall/F1, BLEU/ROUGE-L и др. не применимы. Поэтому нужно перейти от сравни-

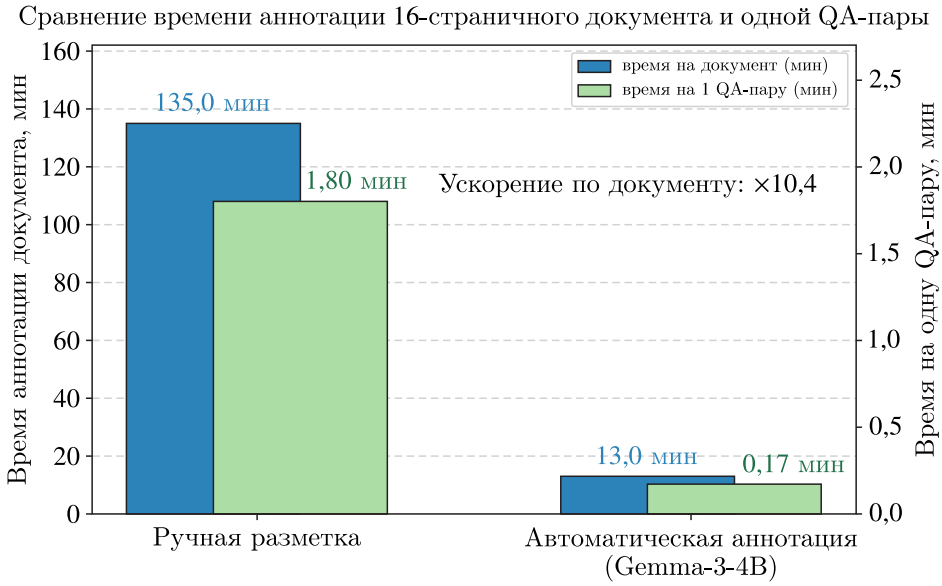

Рис. 2. Сравнение времени аннотации на примере 16-страничного нормативного документа и одной QA-пары при ручной и автоматической разметке (модель Gemma-3-4B)

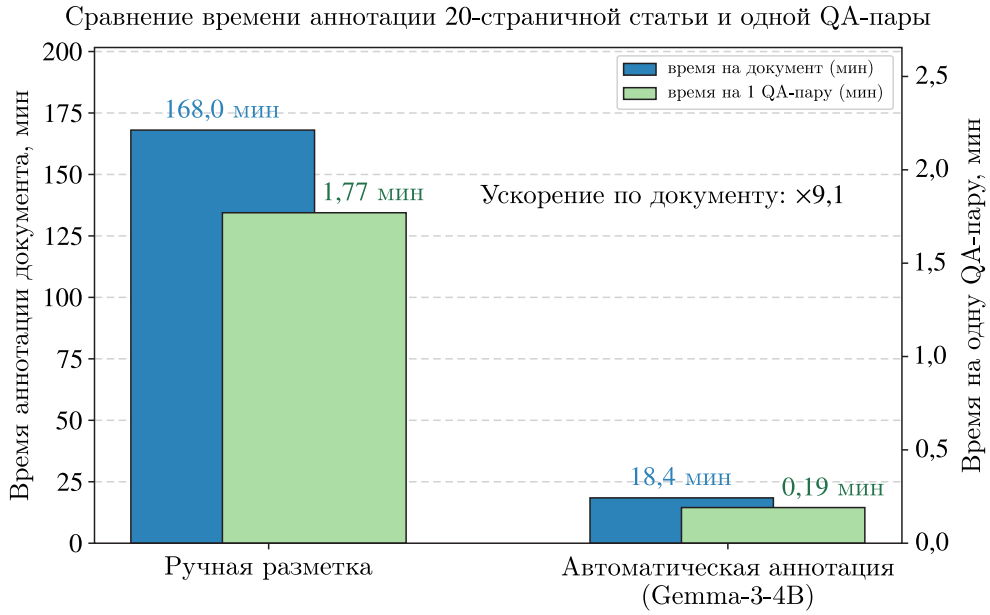

Рис. 3. Сравнение времени аннотации на примере 20-страничной научной статьи и одной QA-пары при ручной и автоматической разметке (модель Gemma-3-4B)

тельного подхода к самооценочному и вероятностному, то есть оценивать качество QA-пар по внутренним признакам согласованности, достоверности и уверенности модели, а не по совпадению с эталоном.

Важнейшая задача заключается в оценке корректности и согласованности автоматически сгенерированных пар, когда отсутствует ручная разметка-эталон, но модель способна сама проверять корректность ответа по исходному контексту или доказательному фрагменту на него опирающегося, либо эту оценку проводит другая языковая модель.

Основную метрику качества принято строить на проверке семантического следования ответа из доказательного фрагмента, а проверку по контексту — использовать как вспомогательную

или диагностическую. При проверке по контексту модель может найти подтверждение в другом месте текста; ошибки в выборе evidence_span маскируются; добавленные факты становятся трудноотличимыми, что снижает строгость оценки.

Так, в данной работе использована метрика семантической согласованности (Contextual Consistency Score, далее — CCS), которая используется как безэталонная мера качества QA-пар и отражает степень семантического следования ответа из соответствующего доказательного фрагмента. Несмотря на использование контекста на этапе генерации вопросов, оценка выполняется исключительно по evidence_span, что обеспечивает более строгую и воспроизводимую проверку фактологической обоснованности ответов.

Метрика основана на идеях *фактической согласованности* и *семантической достоверности*, применяемых в работах по оценке достоверности суммаризации через QA-запросы [Koh et al., 2022; Luo et al., 2024], автоматической проверке *фактической* согласованности (QAFactEval) [Fabbri et al., 2022], разработке метрики выравнивания генерации (AlignScore) [Zha et al., 2023].

Под семантической согласованностью триады «вопрос – ответ – доказательный фрагмент» понимается степень смыслового соответствия, при которой ответ прямо и однозначно следует из доказательного фрагмента и одновременно корректно разрешает семантическое содержание вопроса без привлечения внешних знаний. Если ответ не следует из evidence_span, он некорректен независимо от контекста.

Тем самым CCS выступает безэталонной метрикой внутренней валидации автоматически сгенерированных QA-пар в условиях отсутствия эталонных ответов и позволяет выявлять случаи добавления внешней информации или искажения смысла исходного документа. В отличие от подходов типа QAFactEval, в данной работе предполагается, что контекст соответствует фрагменту исходного документа, вопрос, ответ и доказательный фрагмент генерируются одной моделью, а оценка семантической согласованности выполняется либо той же моделью в режиме самоверификации, либо независимой малой языковой моделью.

Формальное определение

Пусть q_i — сгенерированный моделью вопрос, a_i — сгенерированный моделью ответ, e_i — доказательный фрагмент для i-й QA-пары. Модель-оценщик (верификатор) возвращает скалярную оценку $s_i \in [0, 1]$, характеризующую степень семантического следования ответа a_i из доказательного фрагмента e_i. Тогда

$$\mathrm{CSS}_i = s_i, \quad s_i \in [0, 1].$$

Среднее значение по всем парам:

$$\overline{\mathrm{CSS}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{CSS}_i.$$

При необходимости можно учитывать штраф за противоречие между ответом и evidence_span, что актуально, если оценка проводится той же моделью, которая изначально сгенерировала QA-пары. Модель получает промпт, а затем выдает число, которое имеет соответствующую интерпретацию (табл. 5).

Для оценки QA-пар без эталонного правильного ответа требуются модели, способные:

- сравнивать смысловую согласованность между вопросом, ответом и контекстом,
- давать вероятностную или шкальную оценку,
- быть устойчивыми к переформулировкам.

Таблица 5. Системный запрос и интерпретация семантической согласованности для модели-верификатора

Системный промт	Значение CSS	Интерпретация
Ты — независимый эксперт-оценщик качества разметки для QA-систем. Твоя задача — для каждой записи во входном JSONL строго и воспроизводимо оценить, насколько ответ семантически следует из evidence_span, используя только evidence_span и без внешних знаний. На вход подается многострочный файл в формате JSONL, где каждая строка — независимая запись следующего вида: {"question": "... "answer": "... "evidence_span": "... "type": "basic"}. Поля question и type присутствуют, но не используются при оценке. Каждая строка обрабатывается независимо от других; evidence_span — единственный источник информации. Запрещено использовать: внешние знания, предыдущие или последующие записи, догадки и логические достройки. Если смысл ответа не выражен явно в evidence_span, он считается неподтвержденным. Оценивается только направленное семантическое следование evidence_span → answer. Оцени: подтверждается ли весь смысл ответа доказательным фрагментом, отсутствуют ли в ответе добавленные или уточненные факты. Присвой одно число в диапазоне [0,0; 1,0]. Допускаются промежуточные значения с двумя знаками после запятой. Если ответ добавляет, обобщает или искажает информацию, снизь оценку. Формат вывода: {"CCS": 0.00}	$\geqslant 0{,}85$	Ответ полностью соответствует фрагменту
	0,6–0,85	Частично релевантный ответ, требуется модерация экспертом
	$< 0{,}6$	Ответ не соответствует фрагменту контекста или искажает его

Qwen-2.5-3B обладает именно этими свойствами благодаря унифицированной токенизации и многоязычности, высокой согласованности ответов низкой чувствительности к изменению порядка слов в вопросе. Настройки генерации для ее запуска были практически схожими с параметрами, приведенными выше для модели Gemma-3-4b, за исключением сокращения максимальной длины ответа (до 10 токенов) и установки значения temperature = 0. На один инференсный цикл подавалось 5 QA-пар. В таблице 6 приведены примеры сгенерированных QA-пар моделью Gemma-3-4b и их оценка моделью Qwen-2.5-3b, а также часть контекста, которая непосредственно подтверждает ответ.

Из 1030 проверенных моделью QA-пар 189 получили оценку ниже CSS $\leqslant 0{,}8$, что составляет 18,4 % от общего числа сгенерированных пар; из них 65 пар (6,3 %) были подкорректированы экспертом, а 15 пар (1,5 %) полностью исключены из итогового корпуса.

В табл. 6 под № 2 и № 4 приведены примеры «бракованных» пар с CSS $\leqslant 0{,}8$. В первом случае (пара № 2) числовое значение изменено, добавлен прогнозный характер, которого явно нет в evidence_span (обобщенная интерпретация роста). В втором случае (пара № 4) ответ не воспроизводит явно формулировки из evidence_span (исчезают «конкурсное мероприятие», «потенциал ресурсов», «руководитель и исполнители»); сохраняется общая семантическая связь, но теряется точность соответствия, а также добавляется умеренная степень абстракции, что типично снижает оценку контекстной привязки без перехода в галлюцинацию.

Помимо анализа QA-пар с низкими значениями CCS, была выполнена выборочная проверка пар с высокими значениями метрики (CSS $\geqslant 0{,}85$) на наличие фактических и логических ошибок, включая возможные «галлюцинации». Проверка осуществлялась в форме стратифицированной экспертной валидации, при которой из каждого документа отбиралась фиксированная

Таблица 6. Примеры сгенерированных QA-пар и их оценка моделью Qwen-2.5-3b

№	Примеры QA-пар	Точный фрагмент контекста, из которого взят ответ	CSS
1	{"question": "Для каких целей может быть использовано компьютерное зрение в научно-исследовательской деятельности? "answer": "Распознавание текста на изображениях массива графических материалов, сегментация и идентификация элементов на массиве фотоматериалов, автоматизация вычисления геометрических характеристик элементов."}	{"evidence_span": "Технологии компьютерного зрения могут быть использованы в научно-исследовательской деятельности, в образовании с целью: распознавания текста на изображениях массива графических материалов; идентификации на массиве фотоматериалов и изображений элементов, характеризующих объект исследования; автоматизации вычисления на массиве фотоматериалов и изображений характеристик элементов, характеризующих объекты исследования."}	0,96
2	{"question": Какие тенденции и риски в области кибербезопасности отмечаются в аналитических прогнозах на 2023 год? "answer": "Аналитические оценки указывают на существенное усиление киберугроз, при котором рост активности атак может достигать порядка 40 %, а кибератаки на критическую инфраструктуру рассматриваются как один из приоритетных глобальных рисков, усиливающихся вследствие распространения технологий искусственного интеллекта."}	{"evidence_span": "За 2022 год объем кибератак вырос на 38 %. По прогнозам специалистов, в 2023 году активность хакеров увеличится из-за доступности технологий искусственного интеллекта. Согласно отчету Всемирного экономического форума «Global Risks Perception Survey 2022–2023», в числе главных рисков на 2023 год признаны кибератаки на критическую инфраструктуру."}	**0,60**
3	{"question": "В каких формах может проходить апробация результатов научных исследований? "answer": "Апробация результатов может проходить в форме публичного представления научной и педагогической общественности в форме докладов и выступлений на семинарах, симпозиумах, форумах, конференциях и т. п."}	{"evidence_span": "Апробация результатов может проходить в форме публичного представления научной и педагогической общественности в форме докладов и выступлений на семинарах, симпозиумах, форумах, конференциях и т. п."}	1,0
4	{"question": "Какие задачи может решать технология искусственного интеллекта в области образования? "answer": "Технологии искусственного интеллекта в сфере образования могут использоваться для поддержки управленческих и аналитических решений, включая оценку возможностей участия в различных мероприятиях и оказание помощи в формировании состава участников научной деятельности."}	{"evidence_span": "Технологии искусственного интеллекта могут быть применимы для выполнения следующих задач: – оценка перспективности участия в конкурсном мероприятии с учетом потенциала ресурсов, имеющихся в распоряжении научного коллектива; – рекомендации руководителя и исполнителей научного исследования."}	**0,78**

доля QA-пар с высоким скором. В ходе проверки не выявлено систематических галлюцинаций или фактических искажений, однако зафиксированы единичные случаи упрощения формулировок и частичного опущения контекстных деталей.

Полная ручная валидация всего корпуса QA-пар (золотой стандарт) в рамках настоящего исследования не проводилась в силу высокой трудоемкости и значительных экспертных затрат. Вместе с тем сочетание автоматической оценки семантической согласованности и выборочной

экспертной проверки пар с высокими и низкими значениями CCS позволяет рассматривать полученные оценки качества как обоснованное приближение к полной валидации корпуса.

Следует отметить, что в рамках настоящего исследования процедура автоматической верификации качества QA-пар была ориентирована преимущественно на проверку семантической согласованности связки «evidence_span – ответ». Такой выбор обусловлен инженерной задачей контроля фактологической корректности ответа относительно извлеченного доказательного фрагмента и соответствует распространенной практике безреференсной оценки в условиях отсутствия эталонных ответов. Вместе с тем данный подход не учитывает в явном виде согласованность между формулировкой вопроса и выбранным доказательным фрагментом. Анализ полной триады «вопрос – ответ – evidence_span», включая проверку релевантности evidence_span по отношению к вопросу и выявление случаев латентного несоответствия, рассматривается как отдельное направление дальнейших исследований и требует расширения используемой схемы оценки.

Важно подчеркнуть, что высокие значения CCS не интерпретируются как гарантия отсутствия «галлюцинаций» и искажений, а используются как индикатор вероятной корректности, требующий подтверждения экспертной проверкой на подвыборке данных.

Безусловно, стоит подчеркнуть, что роль эксперта остается ведущей, так как даже при высокой скорости и внешне приемлемом качестве автоматической генерации требуется ручная верификация смысловой точности и полноты ответов. При использовании современных малых языковых моделей остаются случаи неверного интерпретирования смысловых зависимостей, некорректных формулировок или частичных пропусков информации. Вмешательство человека необходимо для финальной калибровки набора данных, согласования формулировок и поддержания высокого уровня валидности корпуса, что особенно важно при подготовке датасета.

Предложенный подход ориентирован на документы с формально-строгой структурой, высокой степенью логической связности и минимальной долей имплицитных знаний (нормативные акты (стандарты), инструкции, научные статьи). Его применение к текстам, включающим сложные расчеты, или слабо структурированным корпусам, требующим внешних знаний, может приводить к снижению качества QA-пар и росту доли экспертной корректировки. Кроме того, подход не предназначен для полностью автоматического формирования золотых стандартов и предполагает обязательный этап человеческой валидации.

Вообще говоря, тема оценки качества генерации QA-пар и апробации различных метрик является довольно обширной и выходит за рамки данной научной статьи. В дальнейшем планируется проведение дополнительного исследования, направленного на анализ применимости и информативности безреференсных метрик для автоматической оценки смысловой связности, адекватности и разнообразия сгенерированных вопросов и ответов без опоры на эталонные данные.

На рис. 4 представлено сравнение совокупного времени аннотации при ручной и автоматической обработке документа серии ГОСТ. Как ранее упоминалось, что использование модели Gemma-3-4B позволило сократить время аннотации с 135 до 13 мин (ускорение $\times$10,4), а при добавлении автоматической оценки качества сгенерированных пар с помощью Qwen-2.5-3B общее время составило 25,5 мин, что по-прежнему обеспечивает более чем пятикратное ускорение по сравнению с ручной разметкой при сохранении приемлемого качества QA-пар. На рис. 5 также приведен замер времени аннотации научной статьи с оценкой качества, аналогично обеспечивающий более чем 4-кратное ускорение времени разметки.

В целом использование модели Qwen-2.5-3B в роли локального верификатора качества QA-пар представляет собой практически обоснованное инженерное решение, ориентированное на полностью офлайн-сценарии и ограниченные вычислительные ресурсы. Вместе с тем следует

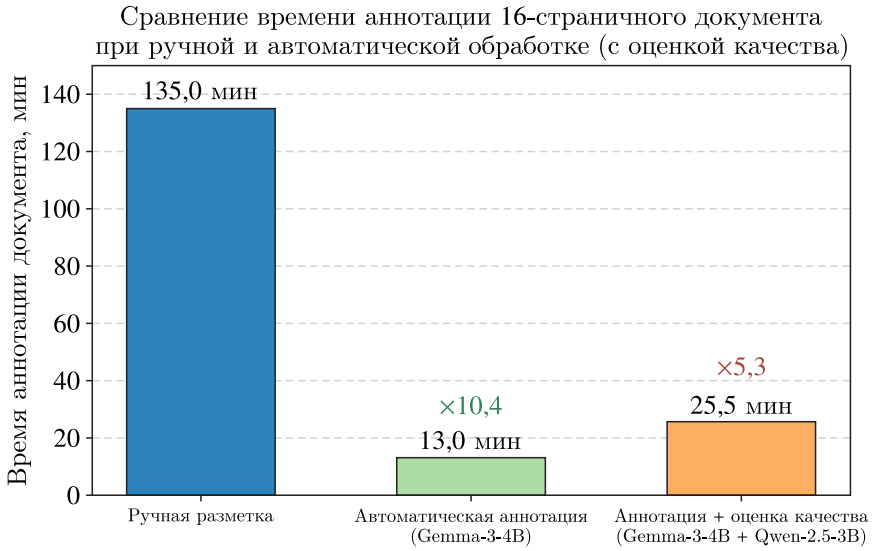

Рис. 4. Сравнение времени аннотации 16-страничного документа при ручной разметке, автоматической генерации (Gemma-3-4B) и аннотации с оценкой качества (Gemma-3-4B и Qwen-2.5-3B)

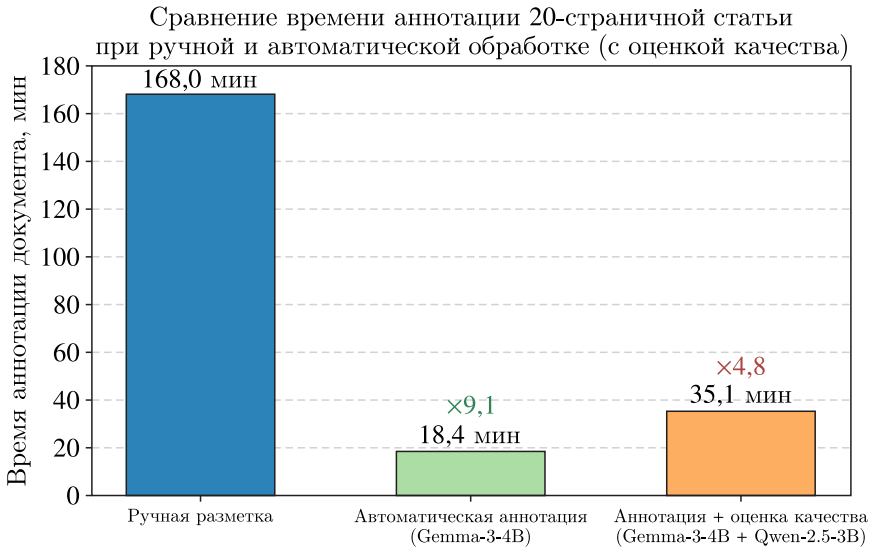

Рис. 5. Сравнение времени аннотации 20-страничной научной статьи при ручной разметке, автоматической генерации (Gemma-3-4B) и аннотации с оценкой качества (Gemma-3-4B и Qwen-2.5-3B)

отметить, что применяемая в настоящей работе метрика CCS отражает внутреннюю оценку языковой модели и на текущем этапе не откалибрована по отношению к внешним экспертным оценкам. Вопросы устойчивости данной оценки, ее корреляции с человеческой экспертизой, а также анализа межмодельной согласованности и статистической стабильности (в том числе с использованием бутстрап-процедур и ансамблей моделей-экспертов) выходят за рамки проведенного эксперимента и рассматриваются как предмет дальнейших исследований.

4. Заключение

Полученные в работе результаты подтверждают возможность полноценного функционирования квантованных языковых моделей на стандартной CPU-платформе без использования GPU, что особенно важно в условиях ограниченных вычислительных ресурсов и требований к защите

данных. Модель Gemma-3-4B является рациональным выбором для задач локальной генерации и аннотации QA-пар, обеспечивая высокую скорость и низкое потребление ресурсов. При этом сохраняется высокий уровень смысловой связности и разнообразия генерируемых ответов, что делает Gemma-3-4B подходящей для использования в автономных программных комплексах, ориентированных на аннотацию и тестирование языковых моделей в защищенных средах.

Модель Qwen-2.5-3B подходит для оценки качества QA-пар, поскольку сочетает в себе достаточную когнитивную сложность для понимания вопросно-контекстных зависимостей; так же как и Gemma-3-4B имеет низкие требования к ресурсам, что позволяет локальное использование, высокую семантическую точность и устойчивость при оценке согласованности ответов.

В отличие от большинства современных работ, ориентированных на применение крупных облачных языковых моделей, предложенный подход нацелен на условия ограниченных вычислительных ресурсов и повышенных требований к автономности и защите данных. Проведенный эксперимент на расширенной выборке документов подтвердил воспроизводимость результатов по производительности и показал, что автоматизированная аннотация обеспечивает устойчивое ускорение по сравнению с ручной разметкой при сохранении необходимости экспертной модерации части сгенерированных данных.

Полученные результаты открывают ряд научных и прикладных перспектив. Во-первых, они создают основу для дальнейших исследований по адаптации и донастройке малых языковых моделей под задачи генерации структурированных обучающих данных. Во-вторых, предложенный подход может быть использован при построении автономных пайплайнов подготовки датасетов для дообучения и оценки языковых моделей в защищенных доменах. В-третьих, дальнейшее расширение экспериментального корпуса и формирование эталонных выборок позволят более строго исследовать корреляцию автоматических метрик с человеческой оценкой и уточнить границы применимости безреференсных методов контроля качества.

Список литературы (References)

Busta L., Oyler A. R. Small language models enable rapid and accurate extraction of structured data from unstructured text: An example with plants and their specialized metabolites // Quantitative Plant Biology. — 2025. — Vol. 6. — P. e26.

Errica F., Sanvito D., Siracusano G., Bifulco R. What did I do wrong? Quantifying LLMs' sensitivity and consistency to prompt engineering // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2025), Volume 1: Long Papers. — Albuquerque, New Mexico: Association for Computational Linguistics, 2025. — P. 1543–1558.

Fabbri A. R., Wu C.-S., Liu W., Xiong C. QAFactEval: improved QA-based factual consistency evaluation for summarization // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2022). — Seattle, United States: Association for Computational Linguistics, 2022. — P. 2587–2601.

Gao M., Hu X., Yin X., Ruan J., Pu X., Wan X. LLM-based NLG evaluation: current status and challenges // Computational Linguistics. — 2025. — Vol. 51, No. 2. — P. 661–687.

Gehrmann S., Clark E., Sellam T. Repairing the cracked foundation: a survey of obstacles in evaluation practices for generated text // Journal of Artificial Intelligence Research. — 2023. — Vol. 77. — P. 103–166.

Jahan M., Wang H., Thebaud T., Sun Y., Le G. H., Fagyal Z., Scharenborg O., Hasegawa Johnson M., Moro Velazquez L., Dehak N. Finding spoken identifications: using GPT-4 annotation for an efficient and fast dataset creation pipeline // Proceedings of the 2024 Joint International Conference

on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). — Torino, Italia: ELRA and ICCL, 2024. — P. 7296–7306.

Kim H., Hwang H., Lee J., Park S., Kim D., Lee T., Yoon C., Sohn J., Park J., Reykhart O., Fetherston T., Choi D., Kwak S. H., Chen Q., Kang J. Small language models learn enhanced reasoning skills from medical textbooks // npj Digital Medicine. — 2025. — Vol. 8. — Article 240.

Koh H. Y., Ju J., Liu M., Pan S. An empirical survey on long document summarization: datasets, models and metrics // ACM Computing Surveys. — 2022. — Vol. 55, No. 8. — P. 1–35.

Luo Z., Xie Q., Ananiadou S. Factual consistency evaluation of summarization in the Era of large language models // Expert Systems with Applications. — 2024. — Vol. 254. — Article 124456.

Ore J., Herzig J., Berant J. An empirical study on type annotations: accuracy, speed, and subjectivity // ACM Transactions on Programming Languages and Systems. — 2021. — Vol. 44, No. 4. — Article 27.

Schroeder H., Roy D., Kabbara J. Just put a human in the loop? Investigating LLM-assisted annotation for subjective tasks // Findings of the Association for Computational Linguistics: ACL 2025. — Vienna, Austria: Association for Computational Linguistics, 2025. — P. 25771–25795.

Shi T., Chen K., Zhao J. Safer-instruct: aligning language models with automated preference data // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1: Long Papers. — Mexico City, Mexico: Association for Computational Linguistics, 2024. — P. 7636–7651.

Sindhujan A., Kanojia D., Orăsan C. Reference-less evaluation of machine translation: navigating through the resource-scarce scenarios // Information. — 2025. — Vol. 16, No. 10. — Article 916.

Son G., Kim M. Multi-level prompting: Enhancing model performance through hierarchical instruction integration // Knowledge-Based Systems. — 2025. — Article 113545.

Xia M., Maharjan S., Le T., Taylor W., Song M. SYNCode: synergistic human – LLM collaboration for enhanced data annotation in stack overflow // Information. — 2025. — Vol. 16, No. 5. — Article 392.

Zha Y., Yang Y., Li R., Hu Zh. AlignScore: evaluating factual consistency with a unified alignment function // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers. — Toronto, Canada: Association for Computational Linguistics, 2023. — P. 11328–11348.

Zhen R., Li J., Ji Y., Yang Z., Liu T., Xia Q., Duan X., Wang Z., Huai B., Zhang M. Taming the titans: a survey of efficient LLM inference serving // Proceedings of the 18th International Natural Language Generation Conference (INLG 2025). — 2025. — P. 522–541.

Zheng Y., Chen Y., Qian B., Shi X., Shu Y., Chen J. A review on edge large language models: design, execution, and applications // ACM Computing Surveys. — 2025. — Vol. 57, No. 8. — Article 209. — P. 1–35.

Zhuang Y., Yu Y., Wang K., Sun H., Zhang C. ToolQA: a dataset for LLM question answering with external tools // Advances in Neural Information Processing Systems (NeurIPS 2023). — 2023. — Vol. 36. — P. 50117–50143.