МИНОБРНАУКИ РОССИИ

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

Кафедра теории функций и стохастического анализа

РАЗРАБОТКА И СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ДЛЯ ДЕТЕКЦИИ AI-СГЕНЕРИРОВАННЫХ РУССКОЯЗЫЧНЫХ НАУЧНЫХ ТЕКСТОВ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 248 группы
направления 09.04.03 — Анализ данных

механико-математического факультета
Субботкина Александра Алексеевича

Научный руководитель
д. ф.-м. н., доцент _____ С. П. Сидоров

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2026

ВВЕДЕНИЕ

В современном информационном пространстве, насыщенном цифровым контентом, задача различения текстов, созданных человеком, от текстов, сгенерированных искусственным интеллектом (ИИ), приобрела критическую важность. Стремительное развитие генеративных моделей, таких как GPT-4, LLaMA, Gemini и Claude, привело к появлению текстов, которые по стилистике, грамматической корректности и смысловой связности практически неотличимы от человеческих.

Интернет и различные инфокоммуникационные технологии с каждым годом занимают всё более значительное место в общественных отношениях и взаимодействиях на всех уровнях. В этой связи информация, распространяемая в сети, зачастую воспринимается человеком как современный аналог энциклопедии или справочника, а также телевидения и газет, которым люди привыкли доверять. Однако существует вероятность, что эта информация не всегда отражает действительность и может быть использована для введения пользователей Интернета в заблуждение, распространения заведомо ложной или подстрекающей информации и в других неправомерных целях.

Объёмы создаваемой текстовой информации за последние десятилетия постоянно увеличиваются, что подтверждается развитием дата-центров, интернет-ресурсов различного назначения, электронного документооборота и других технологий. В то же время создание самих текстов уже не является уникальной прерогативой человека. Специальные алгоритмы и программные средства позволяют генерировать тексты автоматически на основе определённых исходных данных.

Массовое автоматическое создание текстов определённой направленности может быть направлено на пропаганду различных идей, включая социальные, политические и даже преступные, манипуляцию населением или парализацию работы определённых электронных ресурсов. Использование искусственных текстов в виде информационного контента популярных или специально разработанных для этого веб-ресурсов позволяет публиковать и распространять уникальные тексты любого содержания в неограниченном количестве.

Таким образом, учитывая значимость интернет-технологий в жизни человека, злоумышленники могут использовать их в своих неправомерных целях. Указанные угрозы для общества и то, что на сегодняшний день задача автоматического выявления такого контента остаётся нерешённой, обуславливают необходимость создания эффективных методов определения искусственно созданных текстов.

Особую остроту проблема приобретает для русского языка, где количество открытых датасетов и адаптированных методов значительно меньше, чем для английского. Подавляющее большинство исследований и датасетов выполнено для английского языка, что создаёт серьёзную проблему для применения существующих методов к русскоязычным текстам. Основные вызовы включают:

1. Отсутствие стандартизированных бенчмарков для русского языка;
2. Различия в морфологии и синтаксисе;
3. Дефицит русскоязычных LLM с открытым доступом к логитам.

Степень разработанности темы

Среди учёных, добившихся значительных успехов в области текстовой атрибуции и детекции искусственных текстов, можно отметить Д.В. Хмелева, который в своих работах доказал эффективность использования статистического анализа текстовых характеристик. К. Киреев разработал методику, основанную на статистическом анализе текстовых характеристик и текстовых штампов, реализованную в программном средстве «Штампомер». О.Г. Шевелев применил смешанный подход в своём «Стилеанализаторе», используя статистический анализ, деревья решений и нейронные сети. А.С. Романов предложил методику, основанную на использовании опорных векторов в программном средстве «Авторовед».

В последние годы внимание исследователей сосредоточилось на определении авторства коротких текстов, таких как статьи в Интернете, электронные письма, SMS-сообщения и записи в социальных сетях. Особенно активно изучается авторство сообщений электронной почты, что объясняется достаточной длиной этих текстов.

Цель и задачи исследования

Цель работы — разработка и экспериментальное сравнение двух методов детекции AI-сгенерированных текстов: статистического метода GLTR (Giant Language Model Test Room) и нейросетевого метода на основе архитектуры RoBERTa для русскоязычных научных текстов.

Для достижения поставленной цели в работе решаются следующие **задачи**:

1. Провести обзор существующих методов распознавания AI-генерированных текстов, выявить их сильные и слабые стороны.
2. Детально описать современные методы (GLTR, RoBERTa, DetectGPT, водяные знаки) с математической формализацией и анализом ограничений.
3. Адаптировать метод GLTR для русского языка с использованием модели ruGPT-3.5 small.
4. Реализовать и обучить RoBERTa-детектор на русскоязычных данных (датасет AINL-Eval 2025).
5. Провести сравнительный анализ точности и вычислительной эффективности методов GLTR и RoBERTa.
6. Сформулировать практические рекомендации по выбору метода детекции для различных сценариев использования.

Объект и предмет исследования

Объект исследования — научные аннотации, сгенерированные автоматически большими языковыми моделями (LLaMA-3.3-70B, Gemma-2-27B, GPT-4 Turbo), а также исходные естественные тексты.

Предмет исследования — методы GLTR и RoBERTa для классификации текстов по признаку их происхождения (человек / искусственный интеллект).

Научная новизна

Научная новизна работы заключается в следующем:

1. Проведена адаптация метода GLTR для русскоязычных текстов с использованием модели ruGPT-3.5 small.

2. Выполнена экспериментальная валидация RoBERTa-детектора на крупнейшем русскоязычном датасете AINL-Eval 2025 (35 158 текстов в обучающей выборке).
3. Получены количественные оценки сравнительной эффективности статистического и нейросетевого подходов на русскоязычных научных текстах.
4. Выявлены ключевые лексические маркеры, характерные для каждого класса текстов (LLaMA, Gemma, GPT-4 Turbo, human).

Теоретическая и практическая значимость

Теоретическая значимость работы заключается в систематизации знаний о методах детекции AI-сгенерированных текстов и выявлении их применимости к русскоязычным данным.

Практическая значимость результатов исследования состоит в том, что разработанные алгоритмы могут быть использованы:

— в системах модерации пользовательского контента;
— в инструментах проверки академических работ на использование ИИ;
— в исследовательских задачах, связанных с анализом происхождения текстовых материалов;
— как базовые методы сравнения (baseline) при разработке более сложных детекторов.

Методология и методы исследования

В работе использованы следующие методы:

— анализ научной литературы по теме исследования;
— методы статистического анализа текстов (GLTR);
— методы машинного обучения (дообучение трансформерных моделей);
— методы математической статистики (расчёт метрик accuracy, F1-score, AUC-ROC);
— экспериментальное моделирование на реальных данных.

Положения, выносимые на защиту

1. Метод GLTR, адаптированный для русского языка, демонстрирует работоспособность на научных аннотациях, однако его точность (65,83%)

недостаточна для практического применения в ответственных сценариях.

2. Дообученный RoBERTa-детектор на русскоязычных данных достигает точности 85,33%, что на 19,5% выше, чем у GLTR.
3. Выбор метода детекции определяется конкретным сценарием использования: GLTR — для быстрой оценки без обучения, RoBERTa — для максимальной точности при наличии данных и вычислительных ресурсов.

Структура и объём работы

Магистерская работа состоит из введения, пяти разделов, заключения, списка использованных источников (38 наименования) и приложений с программным кодом. Общий объём работы составляет 82 страницы текста. Работа содержит 8 таблиц, 9 рисунков и 2 листинга программного кода.

Первый раздел посвящен систематическому анализу существующих подходов к детекции текстов, порождённых большими языковыми моделями.

Таксономия методов. На основе анализа работ выделены три основных классификационных измерения:

1. **По типу используемых признаков:** статистические, лингвистические, нейросетевые.
2. **По необходимости обучения:** методы, требующие обучения на размеченных данных, zero-shot методы, методы, модифицирующие процесс генерации.
3. **По доступу к генеративной модели:** белые ящики, чёрные ящики, модифицированная генерация.

На пересечении этих измерений выделяются четыре основных класса методов: статистические и лингвистические методы (включая GLTR), методы на основе обученных нейросетевых классификаторов (RoBERTa-детекторы), zero-shot методы (DetectGPT) и технология водяных знаков.

Статистические и лингвистические методы. Проведён детальный анализ метода GLTR, предложенного Геранном и соавторами в 2019 году. Показано, что точность метода на современных LLM падает до 65-70% из-за использования продвинутых стратегий сэмплирования. Рассмотрены лингви-

стические и стилометрические признаки, позволяющие достичь точности до 82% на корпусе научных статей.

Методы на основе машинного обучения. Проанализирована архитектура RoBERTa, её преимущества и ограничения. Показано, что обученные детекторы показывают высокую точность на данных из того же распределения, но демонстрируют катастрофическое падение точности при смене модели-генератора или домена.

Zero-shot методы. Рассмотрен метод DetectGPT, основанный на анализе локальной кривизны поверхности логарифмической вероятности. Выявлены его преимущества (отсутствие обучения, хорошая обобщаемость) и недостатки (высокая вычислительная сложность, необходимость доступа к логарифмам вероятностей).

Атаки и этические аспекты. Проанализирована уязвимость всех постфактум-детекторов к атаке перефразированием (точность падает с 95% до 30-40%). Рассмотрены этические проблемы, связанные с дискриминацией уязвимых групп населения.

Русскоязычные датасеты. Показано, что наиболее значимым русскоязычным датасетом является AINL-Eval 2025, содержащий около 52 000 текстов — аннотаций к научным статьям из 10 доменов, сгенерированных 5 различными LLM.

Второй раздел содержит детальное теоретическое описание четырёх современных методов.

Метод GLTR. Приведена математическая формализация: для текста $x = (x_1, \ldots, x_n)$ вычисляется ранг каждого токена в распределении вероятностей:

$$r_i = |\{v \in \mathcal{V} \mid p_{\mathcal{M}}(v \mid x_{<i}) \geq p_{\mathcal{M}}(x_i \mid x_{<i})\}|$$

Для визуализации ранги группируются в четыре категории с порогами 1, 10, 100. Для автоматической классификации используется метрика «средний логарифмический ранг»:

$$\text{score}_{\text{GLTR}}(x) = \frac{1}{n} \sum_{i=1}^{n} \log(1 - p_{\mathcal{M}}(x_i \mid x_{<i}))$$

Проанализированы ограничения метода: чувствительность к выбору модели, проблемы с короткими текстами, уязвимость к перефразированию.

Метод RoBERTa. Рассмотрена архитектура классификационного варианта RoBERTa. Входная последовательность имеет вид:

$$[[\text{CLS}], x_1, x_2, \ldots, x_n, [\text{SEP}]]$$

Скрытое состояние h0 (соответствует [CLS]-токену) подаётся на классификационную «голову»:

$$\mathbf{z} = \mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \mathbf{h}_0 + \mathbf{b}_1) + \mathbf{b}_2, \quad p(c \mid x) = \frac{e^{z_c}}{\sum_{j=1}^{C} e^{z_j}}$$

Функция потерь — кросс-энтропия. Описаны гиперпараметры обучения: learning rate $2 \cdot 10^{-5}$, batch size 32, количество эпох 3–5.

Метод DetectGPT. Представлена ключевая метрика:

$$\text{score}_{\text{DG}}(x) = \log p_{\mathcal{M}}(x) - \frac{1}{k} \sum_{j=1}^{k} \log p_{\mathcal{M}}(\tilde{x}^{(j)})$$

Описана ускоренная версия FastDetectGPT, использующая аналитическую аппроксимацию кривизны.

Метод водяных знаков. Рассмотрена схема Кирхенбауэра с разделением словаря на «зелёные» и «красные» токены. Приведена формула модифицированного распределения и z-статистика для детекции.

Третий раздел описывает дизайн эксперимента и используемый датасет.

Датасет AINL-Eval 2025. Датасет содержит следующие классы текстов:

Таблица 1 – Структура датасета AINL-Eval 2025

Класс	**Количество в train**
human	8 769
llama-3.3-70b	8 798
gemma-2-27b	8 790
gpt-4-turbo	8 801

Общий объём обучающей выборки составляет 35 158 текстов, тестовой выборки — 6 169 текстов.

Формирование выборки для экспериментов. Для проведения экспериментов с GLTR, учитывая вычислительные затраты на обработку каждого текста языковой моделью, была сформирована сбалансированная выборка объёмом 2 000 текстов (1 000 человеческих и 1 000 сгенерированных AI).

Для RoBERTa данные были разделены на обучающую (1 400 текстов, 70%), валидационную (300 текстов, 15%) и тестовую (300 текстов, 15%) выборки.

Четвертый раздел посвящен эксперименту с методом GLTR.

Выбор модели. Для вычисления вероятностей токенов была выбрана модель `sberbank-ai/rugpt3small_based_on_gpt2` (ruGPT-3.5 small) — русскоязычная версия архитектуры GPT-2, содержащая около 300 миллионов параметров.

Результаты эксперимента. В результате выполнения описанного алгоритма на выборке из 2 000 текстов были получены следующие результаты.

Таблица 2 – Распределение GLTR-метрик

Показатель	**HUMAN**	**AI**	**Разница**
Средняя доля неожиданных токенов	14,80%	11,58%	+3,22%
Медианная доля	14,20%	11,30%	+2,90%
Нормализованный скор	0,3121	0,2818	+0,0303
Средний ранг токенов	81,60	63,50	+18,10

Классификационная точность. При оптимальном пороге классификации threshold = 0,128 были получены метрики:

Таблица 3 – Метрики качества GLTR

Метрика	**Значение**
Общая точность (Accuracy)	65,83%
Распознавание HUMAN (Sensitivity)	67,33%
Распознавание AI (Specificity)	64,33%
F1-score	0,653
AUC-ROC	0,703

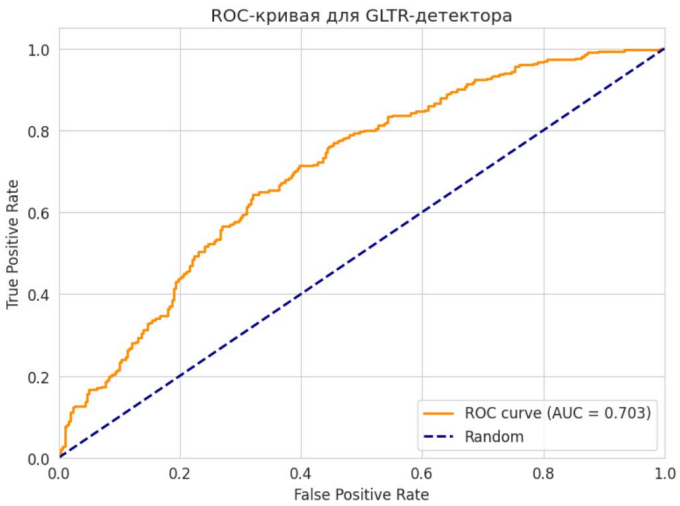

Рисунок 1 – ROC-кривая GLTR-детектора (AUC = 0,703)

Анализ ошибок. Анализ матрицы ошибок показывает, что GLTR допускает примерно одинаковое количество ложноположительных и ложноотрицательных срабатываний.

Таблица 4 – Матрица ошибок GLTR

	Предсказано HUMAN	**Предсказано AI**
Реально HUMAN	67,33%	32,67%
Реально AI	35,67%	64,33%

Пятый раздел посвящен эксперименту с методом RoBERTa.

Параметры эксперимента:

— Модель: `roberta-base-openai-detector`

— Количество эпох: 5

— Размер батча: 8

— Максимальная длина текста: 384 токена

— Скорость обучения: 2×10^{-5}

— Порог уверенности: 0,6

Динамика обучения. В процессе обучения на 5 эпохах наблюдалась устойчивая сходимость модели.

Таблица 5 – Динамика обучения RoBERTa

Эпоха	Train Loss	Train Acc	Train F1	Val Acc	Val F1
1	0,521	78,3%	0,782	74,5%	0,745
2	0,387	84,1%	0,841	79,2%	0,792
3	0,294	88,6%	0,886	82,1%	0,822
4	0,231	91,2%	0,912	83,7%	0,838
5	0,185	93,4%	0,934	84,2%	0,844

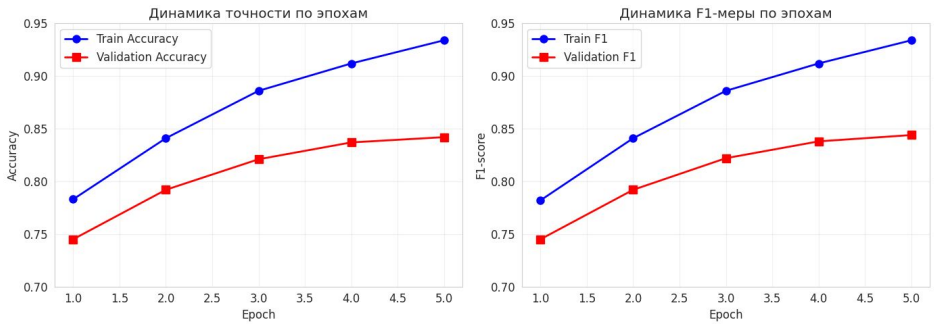

Рисунок 2 – Динамика обучения RoBERTa: точность и F1-мера по эпохам

Результаты на тестовой выборке.

Таблица 6 – Результаты метода RoBERTa на тестовой выборке

Метрика	Значение
Общая точность (Accuracy)	85,33%
F1-score	0,8553
Распознавание HUMAN текстов	84,00% (126/150)
Распознавание AI текстов	86,67% (130/150)
Средняя уверенность правильных ответов	0,87
Средняя уверенность ошибочных ответов	0,58

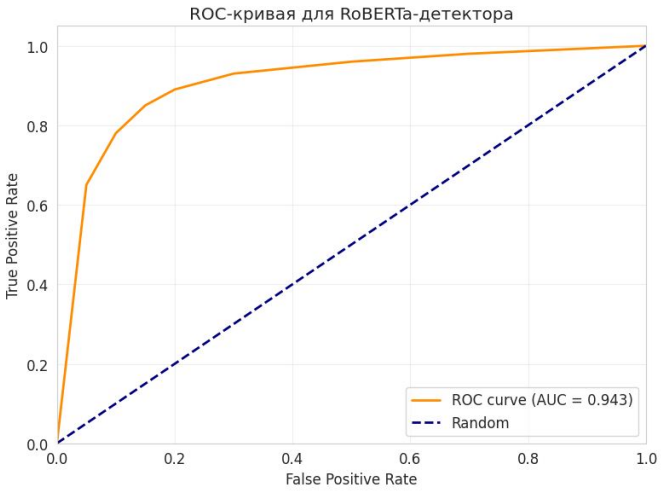

Рисунок 3 – ROC-кривая RoBERTa-детектора (AUC = 0,943)

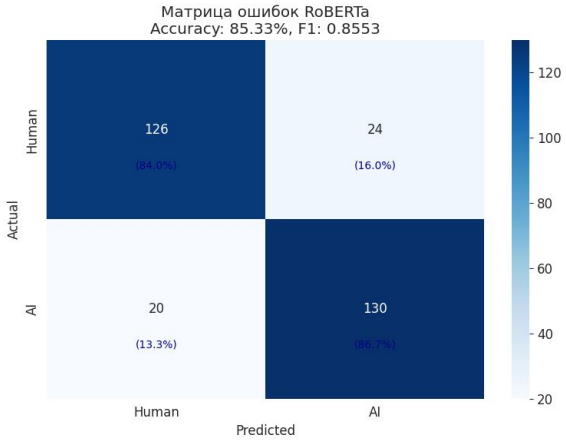

Рисунок 4 – Матрица ошибок RoBERTa-детектора

Сравнительный анализ методов

Таблица 7 – Сравнение методов GLTR и RoBERTa

Критерий	**GLTR**	**RoBERTa**	**Разница**
Общая точность	65,83%	85,33%	+19,50%
Распознавание HUMAN	67,33%	84,00%	+16,67%
Распознавание AI	64,33%	86,67%	+22,34%
Необходимость обучения	Нет	Да	—
Требования к GPU	Опционально	Обязательно	—
Интерпретируемость	Высокая	Низкая	—
Время обработки (2000 текстов)	3 мин	15 мин	—

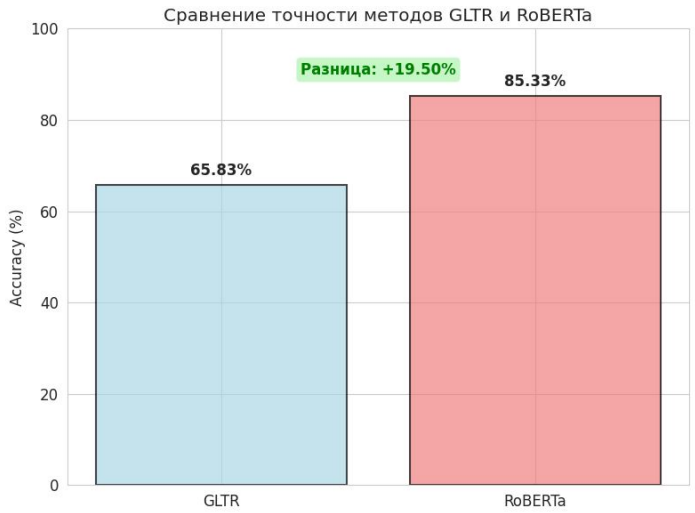

Рисунок 5 – Сравнение точности методов GLTR и RoBERTa

Результаты и выводы

В ходе выполнения диссертационной работы получены следующие основные результаты:

1. Результаты адаптации метода GLTR для русского языка

— Подтверждена основная гипотеза метода GLTR: человеческие тексты содержат на 3,22% больше «неожиданных» токенов (слов, которые языковая модель не ожидала увидеть), чем AI-сгенерированные.
— Средняя доля неожиданных токенов для human составила 14,80%, для AI — 11,58%.
— Нормализованный скор для human — 0,3121, для AI — 0,2818.
— Средний ранг токенов для human — 81,60, для AI — 63,50.
— При оптимальном пороге классификации точность метода GLTR составила 65,83%, что выше случайного угадывания (50%), но недостаточно для практического применения в ответственных сценариях.
— Преимущества GLTR: интерпретируемость, отсутствие необходимости в обучении, работа на стандартном оборудовании.

2. Результаты реализации RoBERTa-детектора

— После дообучения на 1 400 размеченных текстах метод RoBERTa показал точность 85,33% на тестовой выборке из 300 текстов.
— F1-score составил 0,8553, что подтверждает сбалансированность модели по классам.
— Распознавание HUMAN-текстов: 84,00% (126 из 150), распознавание AI-текстов: 86,67% (130 из 150).
— Средняя уверенность модели для правильных ответов (0,87) значительно выше, чем для ошибочных (0,58), что свидетельствует о способности модели оценивать собственную компетентность.
— Наблюдалась устойчивая сходимость в процессе обучения: валидационная точность выросла с 74,5% до 84,2% за 5 эпох, что свидетельствует об отсутствии переобучения.

3. Результаты сравнительного анализа

— RoBERTa превосходит GLTR по общей точности на 19,5 процентных пункта (85,33% против 65,83%).

— Особенно заметно преимущество RoBERTa в распознавании AI-текстов — разница составляет 22,34%.

— Однако это преимущество достигается ценой необходимости в размеченных данных (не менее 1 400 примеров), вычислительных ресурсах (GPU) и времени на обучение.

— GLTR сохраняет преимущества в сценариях, где нет доступа к размеченным данным или требуется быстрая интерпретируемая проверка без использования GPU.

4. Выполнение поставленных задач

В соответствии с целями, сформулированными во Введении, в работе выполнены следующие задачи:

1. Проведён обзор существующих методов распознавания AI-сгенерированных текстов (раздел 1). Установлено, что все постфактум-методы уязвимы к атаке перефразированием, а для русского языка существует лишь один открытый датасет подходящего масштаба — AINL-Eval 2025.
2. Детально описаны современные методы (GLTR, RoBERTa, DetectGPT, водяные знаки) с математической формализацией и анализом ограничений (раздел 2).
3. Реализован алгоритм GLTR для русского языка с использованием модели ruGPT-3.5 small (раздел 4).
4. Реализован и обучен RoBERTa-детектор на датасете AINL-Eval 2025 (раздел 5).
5. Проведено сравнение точности GLTR и RoBERTa. RoBERTa показал результат 85,33%, GLTR — 65,83%.
6. Сформулированы практические рекомендации по выбору метода детекции для различных сценариев использования.

Ограничения исследования

— Эксперименты проводились только на домене «научные аннотации». Обобщаемость результатов на другие домены (новости, социальные медиа, художественная литература) требует дополнительной проверки.
— Для метода RoBERTa использовалась относительно небольшая обучающая выборка (1 400 текстов). При увеличении объёма данных точность может возрасти.
— Оба метода являются постфактум-детекторами и, согласно литературным данным, уязвимы к атаке перефразированием (точность падает до 30–40%).
— Для метода GLTR использовалась модель ruGPT-3.5 small; применение более мощных моделей (ruGPT-3.5 large, YaLM) может повысить точность, но потребует больших вычислительных ресурсов.

Перспективы дальнейших исследований

1. **Апробация на других доменах.** Проведение экспериментов на других русскоязычных датасетах (новостные тексты, социальные медиа, художественная литература) для оценки обобщаемости методов.
2. **Мультиклассовая классификация.** Расширение задачи с бинарной классификации (человек / AI) до мультиклассовой (определение конкретной модели-генератора: LLaMA, Gemma, GPT-4 Turbo и др.).
3. **Повышение точности GLTR.** Применение более мощных русскоязычных моделей (ruGPT-3.5 large, YaLM-100B) для вычисления вероятностей токенов.
4. **Тестирование устойчивости к атакам.** Экспериментальная оценка устойчивости обоих методов к атаке перефразированием с использованием русскоязычных LLM.
5. **Сравнение с zero-shot методами.** Сравнение с современными zero-shot методами типа DetectGPT (при наличии достаточных вычислительных ресурсов).
6. **Разработка ансамблевых методов.** Создание гибридных детекторов, объединяющих преимущества GLTR (интерпретируемость, работа без обучения) и RoBERTa (высокая точность).

Практические рекомендации

На основе проведённого исследования можно рекомендовать:

— Для **быстрой первичной оценки** происхождения текста при отсутствии размеченных данных и вычислительных ресурсов использовать метод **GLTR**. Метод даёт интерпретируемый результат и позволяет эксперту быстро выявить подозрительные участки текста.

— Для **автоматических систем модерации** контента, инструментов проверки академических работ и систем фильтрации спама, где требуется максимальная точность, рекомендуется использовать метод **RoBERTa** при наличии достаточного количества размеченных данных и вычислительных ресурсов (GPU).

— В **ответственных сценариях** (судебная экспертиза, проверка научных работ на плагиат) рекомендуется комбинировать оба метода: GLTR для визуальной интерпретации, RoBERTa для автоматической классификации.