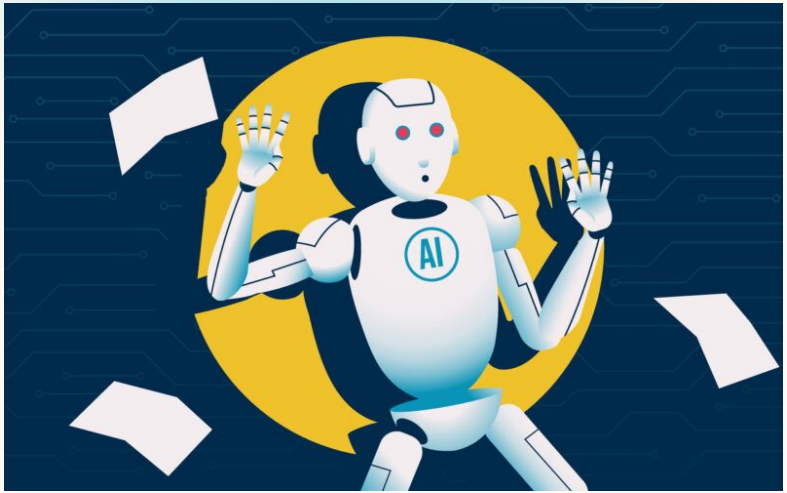
AI-detection для текстов на русском языке



Проект подготовила Алевтина Широкова

Способна ли модель эффективно определять сгенерированные тексты?

Гипотеза:

Да, способна. Но возможно не со 100% эффективностью.

- Антиплагиат, направленный на выявление ИИ
- Интуиция

Цель: Сравнительное исследование различных NLP-подходов для задачи определения сгенерирован ли текст ИИ или написан человеком. Исследовательский интерес не только в итоговом качестве, но и в том, какие методы работают лучше в разных условиях и почему.

На чём училась и тестировалась модель?

Особенности корпуса и датасетов

CoAT🧥 (Corpus of Artificial Texts)

246k human-written texts from publicly available resources

Данные собраны на базе НКРЯ, социальных сетей, Википедии, оцифрованных личных дневников, новостных статей

Artificial texts generated by 13 neural models

Данные собраны на основе выдачи 13 генерирующих моделей, имеющих в инпуте человеческий текст и настроенных для одной или нескольких задач генерации естественного языка: машинный перевод, генерация перефразирования, упрощение текста и обобщение текста.

CoAT🥼 (Corpus of Artificial Texts. Датасеты

На базе CoAT проводились несколько соревнований по двум направлениям:

1. Определение сгенерированных текстов на русском языке (https://www.kaggle.com/competitions/coat-artificial-text-detection), 2024–2025 г.
2. Authorship Attribution среди 14 кандидатов

Датасеты:

train.csv - the training set (contains columns "Id", "Text", "Class", где 0 - H, 1 - M)

```
Train size: 172398
```

val.csv - the validation set (contains columns "Id", "Text", "Class")

```
Validation size: 24628
```

test.csv - the test set (contains columns "Id", "Text")

Выбранные методы

TF-IDF + Logistic Regression (baseline)

- Классический baseline для задачи бинарной классификации текста
- TF-IDF: представление текста как вектора частот слов
- Logistic Regression: линейный классификатор
- Используется как нижняя точка отсчёта качества

Ключевые параметры:

- `max_features = 10000`
- `max_iter = 1000`

Зачем нужен **baseline:**

- Позволяет понять, даёт ли сложная модель реальный прирост
- Быстро обучается, минимальные ресурсы

RuBERT (fine-tuning)

- Предобученная языковая модель для русского языка
- Архитектура BERT
- Дообучение под задачу AI-detection
- Использует контекст и семантику текста

Как обучалась модель:

- Вход: полный текст
- Выход: бинарная метка
  - 0 — текст написан человеком
  - 1 — текст сгенерирован ИИ

Почему RuBERT:

- Оптимизирован под русский язык
- Значительно превосходит классические методы

GPT (zero-shot)

- Большая языковая модель
- Используется без обучения
- Только промпт и текст
- Zero-shot → без примеров в запросе

Формат запроса:

- `"Определи, пожалуйста, был ли этот текст сгенерирован искусственным интеллектом или написан человеком. Ответь 'AI' или 'Human'.\n\nТекст: {text}\nОтвет:"`
- Ответ в виде бинарного класса

DeepSeek (zero-shot)

- Альтернативная LLM
- Тот же zero-shot сценарий
- Те же инструкции и формат ответа

Почему DeepSeek:

- Современная LLM
- Дешевле по токенам
- Интересно сравнить поведение разных моделей

Почему именно эти методы?

1. Классический baseline (TF-IDF + LR)
2. Современная нейросеть (RuBERT)
3. LLM без обучения (GPT, DeepSeek)

Покрытие разных подходов:

- статистический
- нейросетевой
- zero-shot

Такой набор позволяет посмотреть на задачу с разных сторон и понять, где именно появляется прирост качества и за счёт чего.

Результаты

Метрики оценки качества

- **Accuracy** — доля правильных предсказаний
- **F1-score** — баланс **precision** и **recall**
- Основной фокус — **F1-score**
- Оценка проводится на валидационной выборке **(val)**

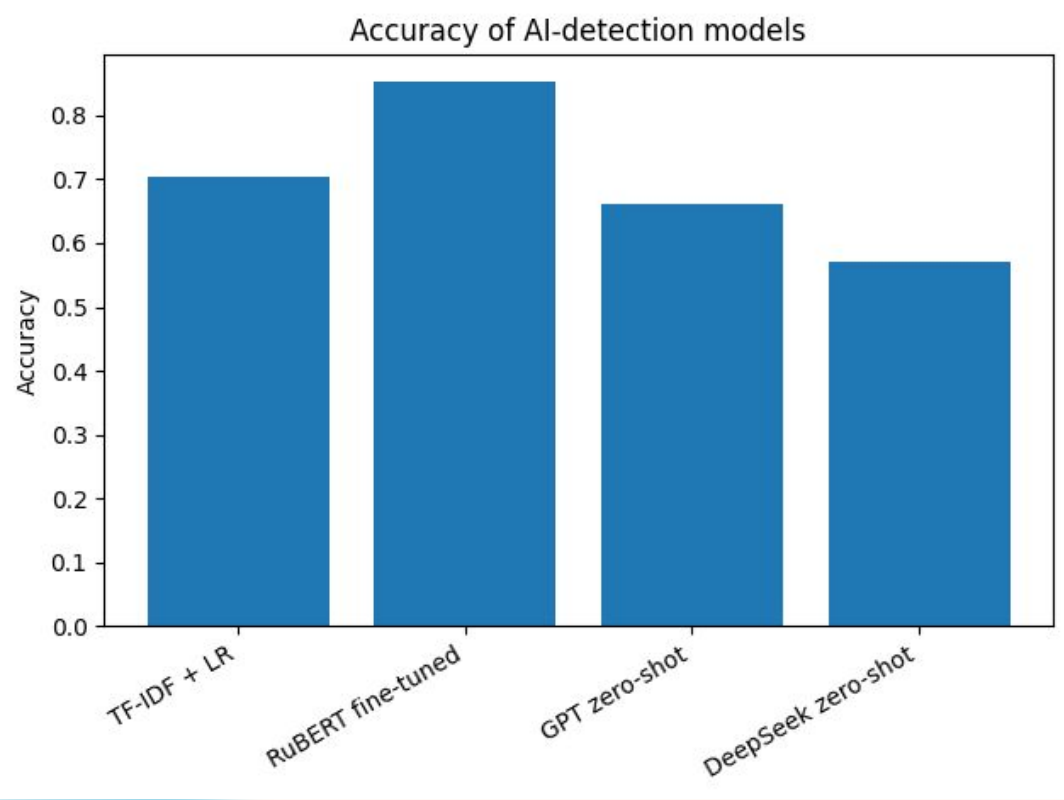
Сравнение моделей по **Accuracy**

TF-IDF + LR: 0.703

RuBERT fine-tuned: 0.852

GPT zero-shot: 0.660

DeepSeek zero-shot: 0.570



Сравнение моделей по **F1-score**

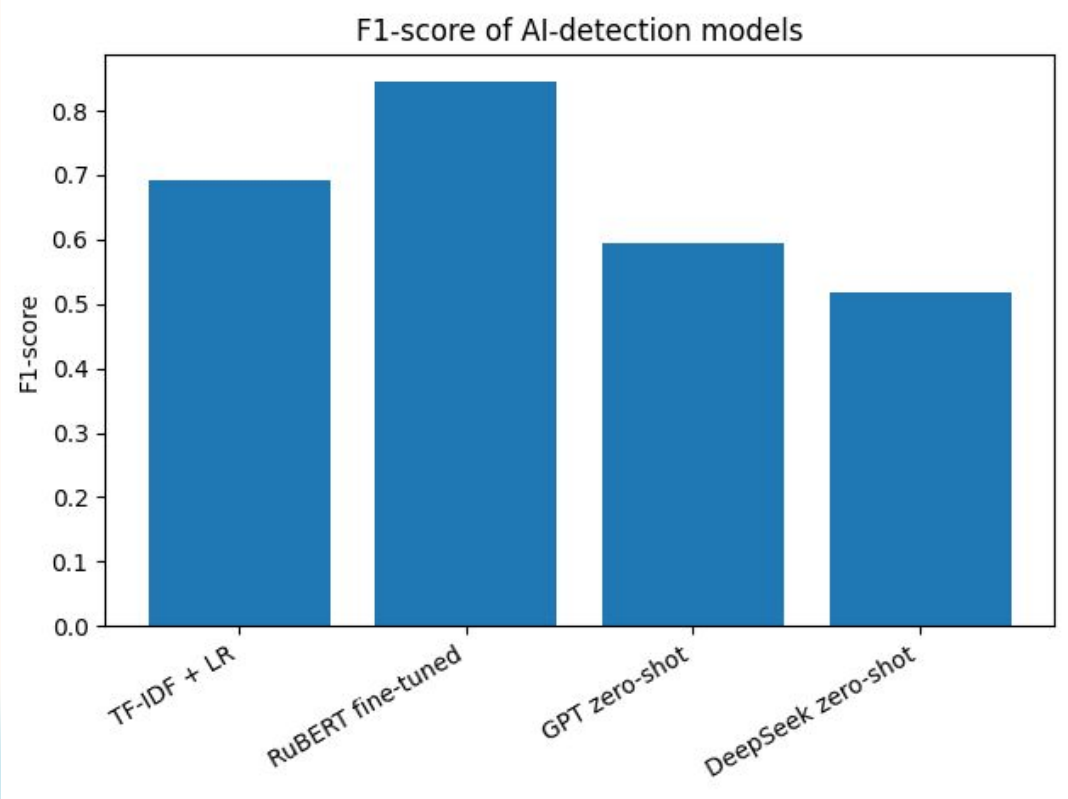
TF-IDF + LR: 0.693

RuBERT fine-tuned: 0.844

GPT zero-shot: 0.595

DeepSeek zero-shot: 0.517

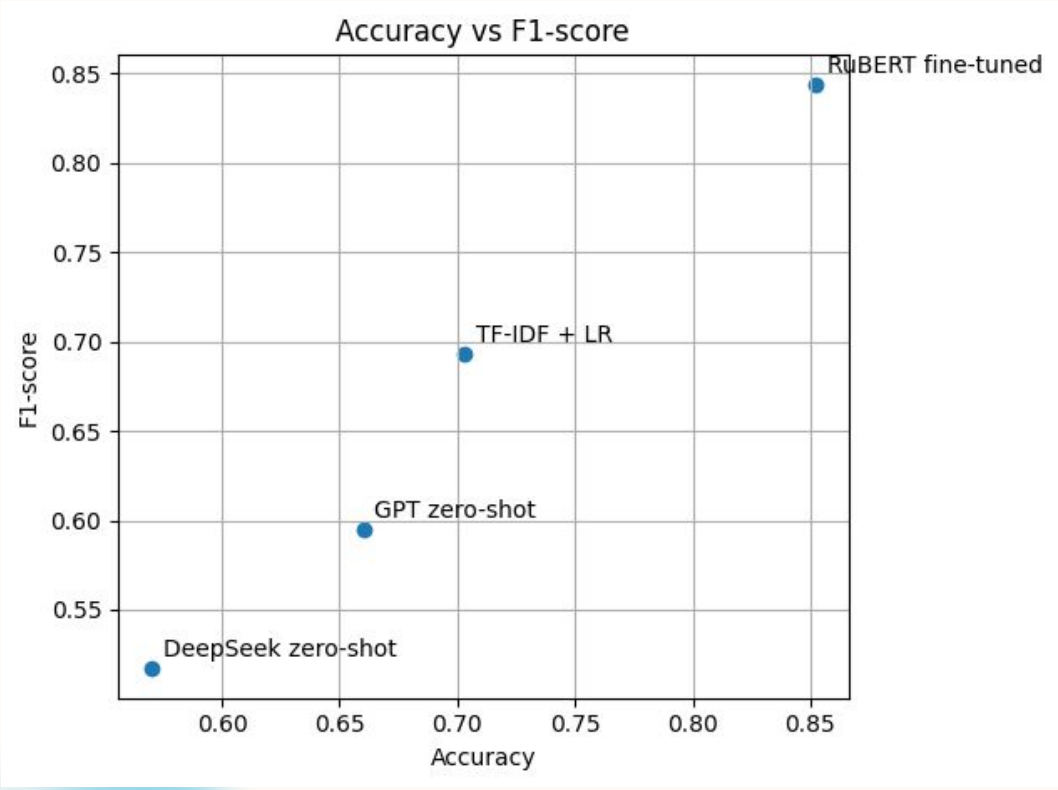
F1-score подчёркивает, что zero-shot модели хуже справляются с балансом ошибок, особенно с определением AI-текстов.



Выводы:

- Модель **RuBERT**, дообученная на целевой задаче, продемонстрировала наивысший **F1-score**, что указывает на её способность эффективно балансировать **precision** и **recall** за счёт контекстных представлений и обучения на размеченных данных.
- Подход **TF-IDF + Logistic Regression** показал стабильный **baseline**-результат, свидетельствующий о наличии различимых лексико-статистических паттернов между человеческими и **AI**-сгенерированными текстами.
- **Zero-shot** большие языковые модели **(GPT** и **DeepSeek)** достигли более низких значений **F1-score** из-за отсутствия адаптации к задаче, что приводит к менее устойчивому разделению классов; при этом **GPT** демонстрирует преимущество над **DeepSeek**.

Fine-tuning является ключевым фактором качества в задачах **AI-detection**



Accuracy vs F1 (scatter plot)

Перспективы

- Улучшение модели RuBERT засчёт экспериментов с гиперпараметрами (таких как размер батча и количество эпох)
- Исследование few-shot стратегий для больших языковых моделей
- Использование альтернативных русскоязычных трансформеров

Спасибо за внимание!