# 

[**ВВЕДЕНИЕ 2**](#_wmwayqnu1zbc)

[Цель практики 2](#_69hl9o1n3duv)

[Задачи практики 2](#_q69i694f2y3n)

[**ОСНОВНАЯ ЧАСТЬ 3**](#_sjpo3myr836v)

[1. Организация выполнения работ 3](#_lb421p96mdzg)

[1.1 Оформление документов на практику 3](#_3sjlk7y9zicx)

[1.2 Ознакомление с требованиями по прохождению практики и правилами оформления отчетности 4](#_rqx6s74bz1y2)

[1.3 Получение индивидуального задания от руководителя практики 4](#_wssi5z69rnkb)

[2. Изучение методики создания наборов данных 5](#_933c76pk1w6l)

[2.1 Проанализировать подходы к формированию вопросов и ответов с ограниченной концептуальной вариативностью 5](#_eq9qgc7tywu1)

[2.2 Изучить алгоритмы генерации контекстно-зависимых данных 6](#_qcvs5h8hnnkc)

[2.3 Выделить ключевые этапы создания датасета (генерация, валидация, тестирование) 6](#_9yqwrj6wt2uv)

[3. Верификация датасета интеллектуальными методами 7](#_wp6jqlnlg0ji)

[3.1 Разработать критерии оценки качества ответов (контекстная согласованность, вариативность) 7](#_ucfbywcopgeu)

[3.2 Предложить методы автоматизированной проверки с использованием БЯМ 8](#_crdxf08yn8o9)

[3.3 Протестировать датасет на примерах, выявить несоответствия 8](#_ublu3bxk9731)

[4. Анализ аномалий в данных 9](#_z7958dr0tv0l)

[4.1 Классифицировать типы аномалий: 9](#_gf85ho8xi8ex)

[4.1.1 Несоответствие ответов контексту 10](#_mgjr7dyxvrkl)

[4.1.2 Дублирование вопросов/ответов 10](#_xz452zzh9ti9)

[4.1.3 Избыточная или недостаточная вариативность 11](#_iyap8ix77dwr)

[4.1.4 Ошибки интерпретации модели 11](#_fxtvyxsbl2fo)

[5. Разработка алгоритмов валидации 11](#_u7wv9vprpk92)

[5.1 Подготовить данные для последующей обработки 11](#_s5opoqi6xxj2)

[5.2 Создать скрипты для автоматического обнаружения аномалий 14](#_cw1qx870qr99)

[5.3 Протестировать алгоритмы на датасете 19](#_ytst0fbssb5t)

[**ЗАКЛЮЧЕНИЕ 22**](#_b85vnsymkdy4)

[Список литературы 23](#_s46o50h9kf43)

# ВВЕДЕНИЕ

Современные большие языковые модели (БЯМ) демонстрируют впечатляющие результаты в генерации текста, однако их предубеждения и контекстная зависимость остаются критическими проблемами.

Для их анализа и устранения требуются специализированные наборы данных, которые позволяют оценить вариативность ответов в зависимости от изменяющихся условий.

Актуальность работы обусловлена необходимостью повышения надежности инструментов оценки БЯМ, особенно в контексте их применения в научных и социально-значимых задачах.

### Цель практики

Провести валидацию и анализ QA-датасета, с использованием интеллектуальных методов для выявления и устранения аномалий.

Цель включает обеспечение качества данных, их соответствия требованиям контекстной зависимости и ограниченной вариативности, а также подготовку рекомендаций для дальнейшего совершенствования методики.

### Задачи практики

1. **Организация выполнения работ**
   1. Оформить документы на практику
   2. Ознакомиться с требованиями по прохождению практики и правилами оформления отчетности
   3. Получить индивидуальное задание от руководителя практики
2. **Изучение методики создания наборов данных**
   1. Проанализировать подходы к формированию вопросов и ответов с ограниченной концептуальной вариативностью
   2. Изучить алгоритмы генерации контекстно-зависимых данных
   3. Выделить ключевые этапы создания датасета (генерация, валидация, тестирование)
3. **Верификация датасета интеллектуальными методами**
   1. Разработать критерии оценки качества ответов (контекстная согласованность, вариативность)
   2. Предложить методы автоматизированной проверки с использованием БЯМ
   3. Протестировать датасет на примерах, выявить несоответствия
4. **Анализ аномалий в данных**
   1. Классифицировать типы аномалий:
      1. Несоответствие ответов контексту
      2. Дублирование вопросов/ответов
      3. Избыточная или недостаточная вариативность
      4. Ошибки интерпретации модели
   2. Подготовить примеры аномалий на основе реальных данных
5. **Разработка алгоритмов валидации**
   1. Подготовить данные для последующей обработки
   2. Создать скрипты для автоматического обнаружения аномалий
   3. Протестировать алгоритмы на датасете

# ОСНОВНАЯ ЧАСТЬ

## 1. Организация выполнения работ

### 1.1 Оформление документов на практику

На начальном этапе выполнены следующие действия:

1. Подписаны договор о практике и дополнительные соглашения с ООО «МЕЖДУНАРОДНЫЙ КОНСОРЦИУМ НАУЧНЫХ КОММУНИКАЦИЙ».
2. Пройдена регистрация в отделе кадров организации, оформлен приказ о допуске к работе.
   1. Изучены локальные нормативные акты компании, включая:
   2. Правила охраны труда и техники безопасности;
   3. Политику информационной безопасности;
   4. Внутренний регламент по работе с практикантами.

### 1.2 Ознакомление с требованиями по прохождению практики и правилами оформления отчетности

Проведен анализ регламентирующих документов:

| **Документ** | **Ключевые требования** |
| --- | --- |
| ГОСТ 7.32-2017 | Объем отчета: 25–35 стр.; структура (титульный лист, введение, аналитическая часть и др.). |
| Методические указания университета | Форматирование: шрифт Times New Roman 14, межстрочный интервал 1.5, поля 30/20 мм. |
| Внутренние стандарты ООО «МЕЖДУНАРОДНЫЙ КОНСОРЦИУМ НАУЧНЫХ КОММУНИКАЦИЙ» | Сроки сдачи: промежуточные материалы - еженедельно, итоговый отчет - до 13.05.2025. |

### 1.3 Получение индивидуального задания от руководителя практики

В ходе встречи с руководителем от Университета (Д.П. Коротков) и куратором от организации (К.В. Короткова) согласованы:

1. **Тема исследования:** «Верификация и валидация наборов данных с контекстно-зависимыми ответами».
2. **План работ:**

| **Этап** | **Сроки** | **Компетенции** |
| --- | --- | --- |
| Изучение методик создания данных | 7–10.04.2025 | ОПК-8, ПК-4 |
| Анализ аномалий | 14–20.04.2025 | ПК-1, ПК-7, УК-1–УК-10 |
| Разработка алгоритмов | 21–27.04.2025 | ПК-6, ПК-8–ПК-10 |

1. **Формат взаимодействия:**

* Онлайн-консультации: 2 раза в неделю;
* Проверка промежуточных результатов: каждую пятницу.

Организационный этап завершен в установленный срок (1 день), что позволило приступить к выполнению основных задач практики.

### 1.4 Итоги раздела

| ПК-7. Способностью использовать отечественные и международные стандарты при проектировании и обеспечении качества прикладного программного обеспечения; | ПК-7.1.  Знает правовые нормы, действующего законодательства, отечественные и международные стандарты в области информационных систем и технологий.  ПК-7.2.  Умеет использовать нормативно-правовые документы, международные и отечественные стандарты при решении стандартных проблем, возникающих в профессиональной деятельности.  ПК-7.3.  Владеет навыками в объеме, позволяющем использовать и составлять нормативно-правовые документы в своей профессиональной деятельности. |
| --- | --- |
| УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач | УК-1.1  Выполняет поиск. необходимой информации, её. критический анализ и обобщает результаты анализа для решения поставленной задачи  УК-1.2  Использует системный подход для решения поставленных задач |

## 2. Изучение методики создания наборов данных

### 2.1 Проанализировать подходы к формированию вопросов и ответов с ограниченной концептуальной вариативностью

Контекстуальная привязка требует учета трех фундаментальных барьеров когнитивных систем:

1. Локальность взаимодействия (пространственно-временные ограничения)
2. Каузальность (причинно-следственные зависимости)
3. Стохастичность (фактор неопределенности)

Моделирование относительных свойств через механизмы селективного внимания, позволяющие выделять системообразующие связи между объектами

Использование концептуальных репрезентаций для ограничения вариативности:

# Псевдокод формирования вопросов

def generate\_question(context, cognitive\_factors):

query = apply\_attention(context, factors=cognitive\_factors)

return add\_constraints(query, allowed\_variants=3)

Введение строгих терминологических дефиниций для ключевых понятий сужает интерпретационное поле вопросов. Например, термин "когнитивная гибкость" определяется как "способность модели адаптировать выводы к изменяющимся контекстам".

### 2.2 Изучить алгоритмы генерации контекстно-зависимых данных

В основном для генерации данных используют многоуровневую архитектуру с использованием:

* Механизмов самореференции (self-attention) для анализа внутренних связей
* Иерархической обработки контекста с разделением на:
  + Локальные паттерны взаимодействия
  + Глобальные системные зависимости

Для динамической адаптации используют следующие шаги:

1. Анализ пространственно-временной локализации
2. Учет каузальных цепочек
3. Моделирование стохастических возмущений

В связи со сложностью системы у данных появляются новые характеристики, которыми не обладают частные участки системы. Такое свойство системы называется “эмерджентность”.

### 2.3 Выделить ключевые этапы создания датасета (генерация, валидация, тестирование)

| **Этап** | **Методы и инструменты** | **Контроль качества** |
| --- | --- | --- |
| Генерация | - Бутстреп-методы с возвратом  - SMOTE-ENC для номинальных признаков  - Контекстно-зависимые трансформеры | Анализ когерентности через механизмы внимания |
| Валидация | - Пилотные эксперименты с тремя БЯМ  - Кросс-валидация по ПКФ-факторам  - Тестирование на парадоксе локальности | Статистика x² для проверки распределений |
| Тестирование | - Многоуровневая оценка эмерджентных свойств  - Анализ устойчивости к стохастическим возмущениям  - Проверка каузальной связности ответов | ROC-анализ для классификаторов предвзятости |

В данной практической работе прикладная часть будет связана с валидацией QA наборов данных.

### 2.4 Итоги раздела

| ПК-9 Способность разрабатывать методы извлечения, анализа и обработки информации | ПК-9.1.  Знать теоретические и прикладные основы анализа данных.  ПК-9.2.  Уметь проводить анализ данных.  ПК-9.3.  Владеть методами извлечения информации и знаний из гетерогенных, мультиструктурированных и неструктурированных источников. |
| --- | --- |
| ПК-10 Способность применять математические методы моделирования процессов обработки информации с использованием средств интеллектуального анализа данных и машинного обучения | ПК-10.1.  Знать принципы решения задач машинного обучения и интеллектуального анализа данных.  ПК-10.2.  Уметь создавать алгоритмические и математические модели прикладных задач интеллектуального анализа данных  ПК-10.3.  Владеть навыками построения описательных и прогнозных аналитических моделей с использованием современных инструментов интеллектуального анализа данных |
| УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач | УК-1.1  Выполняет поиск. необходимой информации, её. критический анализ и обобщает результаты анализа для решения поставленной задачи  УК-1.2  Использует системный подход для решения поставленных задач |
| УК-2  Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся  ресурсов и ограничений | УК-2.1  Формулирует в рамках поставленной цели совокупность задач, обеспечивающих ее достижение  УК-2.2  Выбирает оптимальный способ решения задач, учитывая действующие правовые нормы и имеющиеся условия, ресурсы и. ограничения  УК-2.3  Владеет методиками разработки задач в рамках поставленной цели |

## 3. Верификация датасета интеллектуальными методами

### 3.1 Разработать критерии оценки качества ответов (контекстная согласованность, вариативность)

**Контекстная согласованность** - Нет противоречий между заявлением в контексте и выбранным «target».

**Точность терминологии** - Используются термины, встречающиеся в контексте (не вводятся посторонние понятия).

**Вариативность** - Разные формулировки ответа при идентичном контексте не сказываются на смысле. Модель не копирует «target» дословно, а перефразирует, сохраняя суть.

**Избыточность / лаконичность** - Ответ не содержит лишней информации. Ответ минимален, но достаточен для задачи.

### 3.2 Предложить методы автоматизированной проверки с использованием БЯМ

**Для контекстной согласованности можно использовать следующие решения:**

**NLI (Natural Language Inference)**: Модели типа BERT или RoBERTa для проверки, следует ли ответ из контекста (entailment) или противоречит ему (contradiction).

**Семантическое сходство можно проверить с помощью:**

Сравнения эмбеддингов ответа и контекста через модели типа Sentence-BERT.

Генерация ожидаемого ответа: Запрос к LLM (например, GPT-4) с промптом:

“На основе контекста '{context}' ответь на вопрос '{current\_segment}' одним словом”. Сравнение сгенерированного ответа с target.

**Повышение вариативности:**

**Анализ POS-тегов:** Проверка разнообразия частей речи в ответах (например, не только существительные).

**Кластеризация эмбеддингов:** Группировка ответов с помощью методов UMAP + K-means для выявления семантических повторов.

**Метрики diversity:** Подсчет уникальных лемм, коэффициент Type-Token Ratio (TTR).

### 3.3 Протестировать датасет на примерах, выявить несоответствия

Для примера возьмем 2 сегмента из набора данных и попробуем выявить проблемы:

{

"current\_segment": "Что является ключом к личному успеху?",

"context": "Психолог утверждает, что осознанное принятие неудач и ошибок является важнейшим элементом на пути к успеху, так как именно они закаляют и учат.",

"target": "ошибки"

},

{

"current\_segment": "Что лучше всего мотивирует человека?",

"context": "Социологический эксперимент показал, что страх перед потерей статуса и ресурсов оказывает сильнейшее влияние на мотивацию людей, превышая даже влияние положительного подкрепления.",

"target": "страх"

}

Анализ примеров для удобства оформил в виде таблицы:

| **Сегмент** | **Проблема** | **Рекомендация** |
| --- | --- | --- |
| "Что лучше всего мотивирует человека?" | Контекст упоминает страх потери, но не раскрывает механизм влияния на мотивацию | Добавить каузальные маркеры в контекст |
| "Ключ к личному успеху" | Ответ "ошибки" требует явной связи с концепцией "закаливания" в контексте | Ввести контрольные вопросы для валидации цепочки умозаключений |

3.4 Итоги раздела

| ПК-9 Способность разрабатывать методы извлечения, анализа и обработки информации | ПК-9.1.  Знать теоретические и прикладные основы анализа данных.  ПК-9.2.  Уметь проводить анализ данных.  ПК-9.3.  Владеть методами извлечения информации и знаний из гетерогенных, мультиструктурированных и неструктурированных источников. |
| --- | --- |
| УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач | УК-1.1  Выполняет поиск. необходимой информации, её. критический анализ и обобщает результаты анализа для решения поставленной задачи  УК-1.2  Использует системный подход для решения поставленных задач |

## 4. Анализ аномалий в данных

### 4.1 Классифицировать типы аномалий:

Аномалией считают отклонение данных от ожидаемого или **нормального** поведения. Они могут указывать на ошибки, необычные ситуации или проблемы в системе.

В анализе данных аномалии проявляются как значения или группы значений, значительно отличающиеся от основной массы данных

В нашем случае аномалиями являются нерепрезентативные для оценки модели пары QA вопросов. Рассмотрим аномалии по их типам ниже.

#### 4.1.1 Несоответствие ответов контексту

Это ответ, который не отражают смысл или логику заданного контекста, содержат фактические ошибки или противоречия.

**Примеры:**

* Не связан с ключевой идеей контекста.
* Противоречие между утверждениями в контексте и ответом.

**Методы выявления:**

* Семантический анализ с использованием моделей понимания текста (BERTScore, cosine similarity эмбеддингов).
* Логический анализ причинно-следственных связей.

#### 4.1.2 Дублирование вопросов/ответов

Данная аномалия описывается как повторение идентичных или очень похожих вопросов и/или ответов. Снижает качество и разнообразие датасета.

**Примеры:**

* Повторяющийся вопрос с незначительными изменениями формулировки.
* Повторяющиеся ответы на разные вопросы без контекстуального обоснования.

**Методы выявления:**

* Вычисление текстового сходства (например, Levenshtein distance, Jaccard similarity).
* Кластеризация вопросов и ответов по семантическим признакам.

#### 4.1.3 Избыточная или недостаточная вариативность

**Описание:**

**Избыточная вариативность:** слишком широкий спектр ответов, выходящий за рамки концептуальной ограниченности.

**Недостаточная вариативность:** однообразие ответов, что приводит к переобучению моделей и снижению обобщающей способности.

**Примеры:**

* Много синонимов или близких по смыслу ответов без существенного различия.
* Повторение одного и того же ответа во всех случаях.

**Методы выявления:**

* Анализ лексического разнообразия (уникальные n-граммы, тип-токен соотношение).
* Семантическое кластерирование и распределение по классам

#### 4.1.4 Ошибки интерпретации модели

Ошибки, возникающие из-за неправильной обработки или понимания контекста моделью при генерации ответов.

**Примеры:**

* Модель генерирует ответ, не учитывая важные детали контекста.
* Некорректное связывание причин и следствий.

**Методы выявления:**

* Сравнение ответов, сгенерированных разными моделями.
* Использование логических и контекстных проверок (chain-of-thought reasoning).

### 4.2 Итоги раздела

| ПК-8 Способность выполнять интеллектуальный анализ данных | ПК-8.1.  Знать методы и инструментальные средства интеллектуального анализа данных.  ПК-8.2.  Уметь выбирать средства представления результатов аналитики данных.  ПК-8.3.  Владеть техническими, программными средствами для разработки алгоритмов и программ в области интеллектуального анализа данных |
| --- | --- |
| ПК-9 Способность разрабатывать методы извлечения, анализа и обработки информации | ПК-9.1.  Знать теоретические и прикладные основы анализа данных.  ПК-9.2.  Уметь проводить анализ данных.  ПК-9.3.  Владеть методами извлечения информации и знаний из гетерогенных, мультиструктурированных и неструктурированных источников. |
| УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач | УК-1.1  Выполняет поиск. необходимой информации, её. критический анализ и обобщает результаты анализа для решения поставленной задачи  УК-1.2  Использует системный подход для решения поставленных задач |
| УК-2  Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся  ресурсов и ограничений | УК-2.1  Формулирует в рамках поставленной цели совокупность задач, обеспечивающих ее достижение  УК-2.2  Выбирает оптимальный способ решения задач, учитывая действующие правовые нормы и имеющиеся условия, ресурсы и. ограничения  УК-2.3  Владеет методиками разработки задач в рамках поставленной цели |

## 5. Разработка алгоритмов валидации

Ссылка на git-репозиторий: <https://github.com/xxxNezox/practice-2025>

### 5.1 Подготовить данные для последующей обработки

Подготовка данных для валидации включала несколько шагов, реализованных в Python-скрипте. В первую очередь, необходимо было загрузить данные из файла формата JSONL, каждая строка которого представляет собой отдельную запись с информацией о вопросе, контекстах, ответах и метаданных.

Для удобства работы с данными и структурирования процесса загрузки был разработан класс QADataset, наследующий от torch.utils.data.Dataset. Этот класс выполняет следующие функции:

Чтение файла добавлено конструкторе: класс открывает указанный файл (questions\_data\_en.jsonl), читает его построчно и обрабатывает каждую строку как JSON-объект. Предусмотрена обработка ошибок декодирования JSON и отсутствия файла.

class QADataset(Dataset):

def \_\_init\_\_(self, file\_path: str):

self.data\_entries: List[Dict[str, Any]] = []

self.file\_path = file\_path

try:

with open(self.file\_path, 'r', encoding='utf-8') as f:

for line in f:

if line.strip():

try:

self.data\_entries.append(json.loads(line.strip()))

except json.JSONDecodeError as e:

print(f"Ошибка декодирования JSON в строке: {line.strip()}. Ошибка: {e}")

except FileNotFoundError:

print(f"Ошибка: Файл {self.file\_path} не найден.")

except Exception as e:

print(f"Произошла непредвиденная ошибка при чтении файла {self.file\_path}: {e}")

Метод \_\_getitem\_\_ позволяет получать доступ к отдельным записям по индексу. Внутри этого метода происходит преобразование коротких ключей исходного JSON (например, "q", "v", "c") в более понятные названия ("question", "variations", "general\_context"). Также удаляются поля со значением None, что упрощает дальнейшую обработку.

def \_\_getitem\_\_(self, idx: int) -> Dict[str, Any]:

raw\_item = self.data\_entries[idx]

processed\_item = {

"question": raw\_item.get("q"),

"variations": raw\_item.get("v"),

"field": raw\_item.get("f"),

"subfield": raw\_item.get("s"),

"subject\_matter": raw\_item.get("j"),

"timestamp": raw\_item.get("d"),

"model\_answer\_no\_context": raw\_item.get("n"),

"session\_id": raw\_item.get("t"),

"model\_name": raw\_item.get("model"),

"general\_context": raw\_item.get("c"),

"eval\_answer": raw\_item.get("e"),

"model\_answer\_with\_context": raw\_item.get("m"),

"comparison\_flag": raw\_item.get("k"),

}

return {k: v for k, v in processed\_item.items() if v is not None}

После определения класса, датасет инициализируется с указанием пути к файлу:

file\_path = "/content/drive/MyDrive/pract/questions\_data\_en.jsonl"

dataset = QADataset(file\_path=file\_path)

Таким образом, объект dataset содержит все загруженные и предварительно обработанные записи, готовые для передачи в алгоритмы валидации. Для тестирования алгоритмов на разумном объеме данных выбирается случайная подвыборка записей.

num\_samples\_to\_check = 10

indices\_to\_check = random.sample(range(len(dataset)), k=num\_samples\_to\_check)

Список индексов используется для итерации по выбранным для анализа данным.

В итоге мы получили готовый для дальнейшей работы набор данных, который в дальнейшем можно использовать для валидации.

### 5.2 Создать скрипты для автоматического обнаружения аномалий

Для автоматического обнаружения аномалий были разработаны функции, использующие возможности больших языковых моделей (БЯМ), для оценки качества данных на основе заданных критериев. Основная логика валидации вынесена в две функции, каждая из которых направлена на выявление определенного типа аномалии, описанного в разделе 4.1:

**validate\_answer\_diversity:** Функция проверяет аномалию недостаточной вариативности ответов. Она оценивает, являются ли представленные варианты ответов на один вопрос достаточно разнообразными, или они представляют собой лишь незначительные перефразировки одной и той же идеи, в то время как вопрос мог предполагать более широкий спектр корректных ответов.

Функция принимает на вход элемент данных из датасета. Она извлекает вопрос (question) и список вариантов ответов из поля variations (поле "a" в каждом словаре списка).

Ожидаемый ответ БЯМ- одно слово: "ANOMALY" или "OK". Функция возвращает булево значение и строку с ответом модели.

def validate\_answer\_diversity(item: Dict[str, Any]) -> Tuple[bool, str]:

"""

Аномалия: модель избыточно сужает и отбрасывает варианты.

Проверяет, насколько разнообразны ответы на вопрос, если вопрос подразумевает несколько разных ответов.

"""

question = item.get("question")

variations = item.get("variations") # Список словарей {"c": context, "a": answer}

if not question or not variations or len(variations) < 2: # Нужно хотя бы 2 ответа для сравнения разнообразия

return False, "Недостаточно данных для проверки (нужен вопрос и минимум 2 варианта ответа)"

answers = [var.get("a") for var in variations if var.get("a")]

if len(answers) < 2:

return False, "Недостаточно ответов в 'variations' для проверки разнообразия."

answers\_formatted = "\n".join([f"- {ans}" for ans in answers])

prompt = f"""

Ты — эксперт по оценке качества QA-датасетов.

Твоя задача — оценить РАЗНООБРАЗИЕ ПРЕДОСТАВЛЕННОГО СПИСКА ОТВЕТОВ на заданный вопрос.

Вопрос: "{question}"

Предоставленный список ответов для оценки разнообразия:

{answers\_formatted}

Задание: Проанализируй, являются ли ответы в ПРЕДОСТАВЛЕННОМ СПИСКЕ достаточно разнообразными друг относительно друга.

Если вопрос по своей природе может допускать несколько различных категорий или типов правильных ответов,

а ответы в ПРЕДОСТАВЛЕННОМ СПИСКЕ являются очень похожими, синонимичными, или лишь незначительными перефразировками одной и той же основной идеи,

И ПРИ ЭТОМ НЕ ОХВАТЫВАЮТ потенциальную широту возможных правильных ответов, которую мог бы предполагать вопрос, то это аномалия (недостаток разнообразия).

ВАЖНО: Оценивай разнообразие именно тех ответов, которые даны в "Предоставленном списке ответов".

Даже если вопрос начинался со слов "Назови одну причину...", а в списке дано несколько причин, твоя задача — оценить, отличаются ли эти причины друг от друга.

Пример аномалии (недостаток разнообразия в списке):

Вопрос: "Какие есть популярные виды спорта?"

Предоставленный список ответов:

- Футбол

- Европейский футбол

- Соккер

(Это ANOMALY, так как все ответы в списке про одно и то же, хотя вопрос мог бы предполагать баскетбол, хоккей и т.д.)

Пример отсутствия аномалии (ответы в списке разнообразны):

Вопрос: "Назови одну причину проблем с экологией."

Предоставленный список ответов:

- Вырубка лесов

- Загрязнение океана пластиком

(Это OK, так как "Вырубка лесов" и "Загрязнение океана пластиком" — это РАЗНЫЕ причины, представленные в списке).

Является ли это аномалией (недостаток разнообразия в ПРЕДОСТАВЛЕННОМ СПИСКЕ ОТВЕТОВ)?

Ответь ОДНИМ СЛОВОМ: 'ANOMALY' или 'OK'.

"""

response\_text = call\_gemini\_api(prompt)

if response\_text and "ANOMALY" in response\_text.upper():

return True, f"Обнаружена аномалия недостатка разнообразия. Ответ LLM: {response\_text}"

elif response\_text and "OK" in response\_text.upper():

return False, f"Аномалия недостатка разнообразия не обнаружена. Ответ LLM: {response\_text}"

else:

return False, f"Не удалось определить аномалию недостатка разнообразия. Ответ LLM: {response\_text}"

**validate\_answer\_uniqueness\_in\_context:** Валидирует аномалии множественности ответов в рамках одного конкретного контекста. Проверяется пара "вопрос + контекст" из списка variations и определяется, является ли предложенный ответ единственно верным (или очень близким его перефразированием) в данном контексте, или же существует(ют) другое(ие) существенно отличающиеся верные альтернативы.

def validate\_answer\_diversity(item: Dict[str, Any]) -> Tuple[bool, str]:

"""

Аномалия: модель избыточно сужает и отбрасывает варианты,

разные пары "контекст + ответ" ведут к синонимичным по сути вариантам ответа,

"""

question = item.get("question")

variations = item.get("v") # {"c": context, "a": answer}

if not question or not variations:

return False, "Недостаточно данных: отсутствует вопрос или список вариаций (поле 'v')."

context\_answer\_pairs = []

for var in variations:

context = var.get("c")

answer = var.get("a")

if answer is not None:

display\_context = context if context and str(context).strip() else "Общий/не указан"

context\_answer\_pairs.append({"context": display\_context, "answer": str(answer)})

if len(context\_answer\_pairs) < 2:

return False, "Недостаточно валидных пар 'контекст-ответ' (минимум 2) для проверки дифференциации."

pairs\_formatted\_for\_prompt = []

for i, pair in enumerate(context\_answer\_pairs):

pairs\_formatted\_for\_prompt.append(

f"Пара {i+1}:\n Контекст: \"{pair['context']}\"\n Ответ: \"{pair['answer']}\""

)

formatted\_pairs\_string = "\n\n".join(pairs\_formatted\_for\_prompt)

prompt = f"""

Ты — эксперт по оценке качества QA-датасетов.

Твоя задача — оценить, приводят ли РАЗНЫЕ КОНТЕКСТЫ к СЕМАНТИЧЕСКИ РАЗНЫМ ответам на один и тот же вопрос.

Аномалия ("ANOMALY") возникает, если модель для различных предоставленных контекстов даёт ответы,

которые по своей сути являются синонимами, очень близкими по смыслу или лишь незначительными перефразировками одной и той же идеи,

В ТО ВРЕМЯ КАК РАЗЛИЧНЫЕ КОНТЕКСТЫ ДОЛЖНЫ БЫЛИ БЫ ПРИВЕСТИ К БОЛЕЕ СУЩЕСТВЕННО ОТЛИЧАЮЩИМСЯ АСПЕКТАМ ОТВЕТА ИЛИ РАЗНЫМ ОТВЕТАМ ПО СМЫСЛУ.

То есть, модель не смогла достаточно дифференцировать свои ответы на основе предоставленных различных контекстов, хотя должна была.

Если ответы действительно семантически разные и адекватно отражают нюансы, привнесенные разными контекстами, то это "OK".

Проанализируй следующий вопрос и предоставленные пары "контекст-ответ":

Вопрос: "{question}"

Предоставленные пары "контекст-ответ":

{formatted\_pairs\_string}

Задание:

Оцени, являются ли ответы в этих парах избыточно похожими/синонимичными по СМЫСЛУ, УЧИТЫВАЯ ИХ КОНТЕКСТЫ.

Пример АНОМАЛИИ (один из контекстов общий, другой специфичный, но ответы почти идентичны):

Вопрос: "Каковы преимущества использования солнечной энергии?"

Предоставленные пары "контекст-ответ":

Пара 1:

Контекст: "Общий/не указан"

Ответ: "Снижение выбросов углекислого газа"

Пара 2:

Контекст: "Для частных домовладений"

Ответ: "Уменьшение углеродного следа домохозяйства"

(Это ANOMALY, так как второй ответ является лишь конкретизацией первого без добавления нового смысла, специфичного для частных домовладений, например, "снижение счетов за электричество" или "энергетическая независимость".)

Пример ОТСУТСТВИЯ аномалии (контексты близкие, ответы ожидаемо близкие):

Вопрос: "Как приготовить кофе?"

Предоставленные пары "контекст-ответ":

Пара 1:

Контекст: "Используя турку"

Ответ: "Залить молотый кофе водой в турке и довести до кипения, не давая убежать."

Пара 2:

Контекст: "В капсульной кофемашине"

Ответ: "Поместить капсулу в кофемашину, подождать приготовления"

(Это OK, так как ответы на разные контексты принципиально разные.)

Является ли это аномалией (недостаток смысловой дифференциации ответов при разных контекстах)?

Ответь ОДНИМ СЛОВОМ: 'ANOMALY' или 'OK'.

"""

response\_text = call\_gemini\_api(prompt)

processed\_response = response\_text.strip().upper() if response\_text else ""

if processed\_response == "ANOMALY":

return True, f"Обнаружена аномалия недостатка смысловой дифференциации ответов. Ответ LLM: {response\_text}"

elif processed\_response == "OK":

return False, f"Аномалия недостатка смысловой дифференциации ответов не обнаружена. Ответ LLM: {response\_text}"

else:

return False, f"Не удалось однозначно определить аномалию. Ответ LLM: '{response\_text}'. Ожидался 'ANOMALY' или 'OK'."

Таким образом, разработанные алгоритмы валидации предоставляют способ проверки двух конкретных типов аномалий в QA-датасете, используя возможности БЯМ для семантического и контекстного анализа.

### 5.3 Протестировать алгоритмы на датасете

Тестирование разработанных алгоритмов валидации было проведено на подвыборке загруженного датасета, состоящей из 10 случайных записей. Целью тестирования являлась демонстрация работоспособности скриптов обнаружения аномалий и оценка типа информации, которую они предоставляют.

Для каждой записи последовательно вызывались функции validate\_answer\_diversity и validate\_answer\_uniqueness\_in\_context.

Результаты каждого вызова, включая флаг аномалии и текстовое объяснение от модели, собирались в соответствующие списки diversity\_results\_data и uniqueness\_results\_data.

**Анализ результатов валидации разнообразия ответов:**

| **dataset\_idx** | **question** | **llm\_reason** | **answers\_evaluated** | **is\_anomaly** |
| --- | --- | --- | --- | --- |
| 4691 | What is a common feature of GMO labeling? | Аномалия недостатка разнообразия не обнаружена. Ответ LLM: OK | ['Optional', 'Required'] | FALSE |
| 4768 | What is the main method of GMO regulation in the EU? | Обнаружена аномалия недостатка разнообразия. Ответ LLM: ANOMALY | ['Strict', 'Flexible'] | TRUE |
| 2021 | Which service is known for video conferencing? | Аномалия недостатка разнообразия не обнаружена. Ответ LLM: OK | ['Zoom', 'Google Meet'] | FALSE |
| 443 | Is the coastline paradox related to fractals? | Обнаружена аномалия недостатка разнообразия. Ответ LLM: ANOMALY | ['Yes', 'Yes'] | TRUE |
| 2599 | What is the role of symmetry in determining crystal properties? | Обнаружена аномалия недостатка разнообразия. Ответ LLM: ANOMALY | ['Simple', 'Complex'] | TRUE |
| 4059 | Which habitat type is highly vulnerable to climate change? | Аномалия недостатка разнообразия не обнаружена. Ответ LLM: OK | ['Mangroves', 'Alpine zones'] | FALSE |
| 694 | What is the product of the first 3 positive integers? | Обнаружена аномалия недостатка разнообразия. Ответ LLM: ANOMALY | ['6', '6'] | TRUE |
| 4088 | What is the effect of ocean acidification on marine chemical reactions? | Аномалия недостатка разнообразия не обнаружена. Ответ LLM: OK | ['Decrease', 'Increase'] | FALSE |
| 2761 | Which type of seismic wave arrives first at a seismograph? | Аномалия недостатка разнообразия не обнаружена. Ответ LLM: OK | ['P-wave', 'Rayleigh'] | FALSE |
| 1997 | Which protocol is primarily used for secure web browsing? | Обнаружена аномалия недостатка разнообразия. Ответ LLM: ANOMALY | ['HTTPS', 'HTTP'] | TRUE |

Результаты показывают, что модель успешно выявляет случаи полного дублирования ответов, а также может обнаруживать ситуации, когда варианты ответов, хотя и формально разные, недостаточно полно охватывают концептуальное пространство, подразумеваемое вопросом, что соответствует аномалии недостаточной вариативности.

**Анализ результатов валидации множественности ответов в контексте:**

| **dataset\_idx** | **variation\_idx** | **question** | **context** | **answer** | **llm\_reason** | **is\_anomaly** |
| --- | --- | --- | --- | --- | --- | --- |
| 4691 | 0 | What is a common feature of GMO labeling? | Voluntary | Optional | Аномалия множественности ответов не обнаружена. Ответ LLM: OK | FALSE |
| 4691 | 1 | What is a common feature of GMO labeling? | Mandatory | Required | Аномалия множественности ответов не обнаружена. Ответ LLM: OK | FALSE |
| 4768 | 0 | What is the main method of GMO regulation in the EU? | Precautionary principle | Strict | Обнаружена аномалия множественности ответов. Ответ LLM: ANOMALY | TRUE |
| 4768 | 1 | What is the main method of GMO regulation in the EU? | Risk assessment | Flexible | Обнаружена аномалия множественности ответов. Ответ LLM: ANOMALY | TRUE |
| 2021 | 0 | Which service is known for video conferencing? | Business | Zoom | Аномалия множественности ответов не обнаружена. Ответ LLM: OK | FALSE |

Было выявлено, что для данных пар "Вопрос + Контекст" БЯМ посчитала, что существуют другие, существенно отличающиеся правильные ответы, помимо того, который был предложен в датасете.

### 5.4 Итоги раздела

| ПК-1. Способность разработки прикладного программного обеспечения, автоматизации работы с базами данных и документами, программирования бизнес-логики приложений, интеграции разнородных данных | ПК-1.1.  Знать технологии программирования прикладного программного обеспечения и бизнес-логики приложений.  ПК-1.2.  Уметь разрабатывать и конфигурировать прикладное программное обеспечение.  ПК-1.3.  Владеть навыками автоматизации решения типовых задач, работы с базами данных и документами, интеграции разнородных данных в корпоративных информационных системах. |
| --- | --- |
| ПК-8 Способность выполнять интеллектуальный анализ данных | ПК-8.1.  Знать методы и инструментальные средства интеллектуального анализа данных.  ПК-8.2.  Уметь выбирать средства представления результатов аналитики данных.  ПК-8.3.  Владеть техническими, программными средствами для разработки алгоритмов и программ в области интеллектуального анализа данных |
| ПК-9 Способность разрабатывать методы извлечения, анализа и обработки информации | ПК-9.1.  Знать теоретические и прикладные основы анализа данных.  ПК-9.2.  Уметь проводить анализ данных.  ПК-9.3.  Владеть методами извлечения информации и знаний из гетерогенных, мультиструктурированных и неструктурированных источников. |
| ПК-10 Способность применять математические методы моделирования процессов обработки информации с использованием средств интеллектуального анализа данных и машинного обучения | ПК-10.1.  Знать принципы решения задач машинного обучения и интеллектуального анализа данных.  ПК-10.2.  Уметь создавать алгоритмические и математические модели прикладных задач интеллектуального анализа данных  ПК-10.3.  Владеть навыками построения описательных и прогнозных аналитических моделей с использованием современных инструментов интеллектуального анализа данных |
| УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач | УК-1.1  Выполняет поиск. необходимой информации, её. критический анализ и обобщает результаты анализа для решения поставленной задачи  УК-1.2  Использует системный подход для решения поставленных задач |
| УК-2  Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся  ресурсов и ограничений | УК-2.1  Формулирует в рамках поставленной цели совокупность задач, обеспечивающих ее достижение  УК-2.2  Выбирает оптимальный способ решения задач, учитывая действующие правовые нормы и имеющиеся условия, ресурсы и. ограничения  УК-2.3  Владеет методиками разработки задач в рамках поставленной цели |

# ЗАКЛЮЧЕНИЕ

В ходе выполнения работы были последовательно решены поставленные задачи. Изучены подходы к формированию QA-пар с учетом концептуальной вариативности и контекстной зависимости, проанализированы этапы создания датасетов. Ключевым аспектом стала разработка критериев оценки качества ответов (контекстная согласованность, вариативность) и предложение методов автоматизированной проверки с привлечением больших языковых моделей (БЯМ).

Тестирование разработанных алгоритмов на случайной подвыборке датасета продемонстрировало их работоспособность и способность успешно выявлять потенциальные аномалии. В частности, были обнаружены записи, где представленные варианты ответов были слишком похожи или являлись полными дубликатами (аномалия недостатка разнообразия).

Таким образом, цель практики достигнута. Разработанные интеллектуальные методы валидации позволяют автоматизировать процесс проверки QA-датасетов, повышая их качество и надежность для использования в задачах оценки БЯМ.

## Список литературы

1. Vaswani A., Shazeer N., Parmar N. и др. Attention is all you need // URL: <https://arxiv.org/pdf/1706.03762.pdf> (дата обращения: 17.05.2025).
2. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // URL: <https://arxiv.org/pdf/1810.04805.pdf> (дата обращения: 17.05.2025).
3. Reimers N., Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks // URL: <https://arxiv.org/pdf/1908.10084.pdf> (дата обращения: 17.05.2025).
4. McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction // URL: <https://arxiv.org/pdf/1802.03426.pdf> (дата обращения: 17.05.2025).
5. Radford A., et al. GPT-4 Technical Report // URL: <https://cdn.openai.com/papers/gpt-4.pdf> (дата обращения: 17.05.2025).
6. Liu Y., Ott M., Goyal N. и др. RoBERTa: A Robustly Optimized BERT Pretraining Approach // URL: <https://arxiv.org/pdf/1907.11692.pdf> (дата обращения: 17.05.2025).
7. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey // URL: <https://dl.acm.org/doi/10.1145/1541880.1541882> (дата обращения: 17.05.2025).
8. Документация PyTorch: torch.utils.data.Dataset // URL: <https://pytorch.org/docs/stable/data.html> (дата обращения: 17.05.2025).