**НИТУ «МИСИС»**

**Кафедра инженерной кибернетики**

**ОТЧЕТ**

по

**ЛАБОРАТОРНОЙ РАБОТЕ №4**

*«Разработка асинхронного чат-бота для мессенджера Telegram*

*с использованием языковых моделей архитектур GPT и LLaMA»*

учебная дисциплина

«Методы искусственного интеллекта»

**Студент: Осипова Елизавета Андреевна**

**Группа**: \_\_\_\_\_**БПМ-21-2**\_\_\_\_\_\_\_

**Преподаватель: Хонер П. Д.**

**Оценка: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Дата защиты: \_\_\_\_\_\_\_\_\_\_\_\_\_**

**2024 г.**

Оглавление

[**ВВЕДЕНИЕ** 3](#_Toc191395958)

[**1.** **Задание на лабораторную работу** 4](#_Toc191395959)

[**1.1. Основная цель работы** 4](#_Toc191395960)

[**1.2. Общие сведения** 4](#_Toc191395961)

[**1.3. Языковые модели, рекомендуемые к использованию** 4](#_Toc191395962)

[**2.** **Выполнение лабораторной работы** 5](#_Toc191395963)

[**2.1. Выбор библиотеки для создания чат-бота** 5](#_Toc191395964)

[**2.2. Выбор предметной области** 5](#_Toc191395965)

[**2.3. Проектирование чат-бота** 5](#_Toc191395966)

[**2.4. Разработка асинхронного чат-бота** 6](#_Toc191395967)

[**2.5. Использованные технологии и библиотеки** 6](#_Toc191395968)

[**2.6. Сравнительный анализ языковых моделей** 6](#_Toc191395969)

[**2.7. Проблемы и ограничения** 7](#_Toc191395970)

[**2.8. Модулярность кода** 7](#_Toc191395971)

[**2.9. Использование файла `educational\_keywords.txt`** 10](#_Toc191395972)

[**ВЫВОДЫ** 12](#_Toc191395973)

[**ИСПОЛЬЗУЕМАЯ ЛИТЕРАТУРА** 13](#_Toc191395974)

[**ПРИЛОЖЕНИЕ А. Работа чат-бота** 14](#_Toc191395975)

[**ПРИЛОЖЕНИЕ Б. Код чат-бота** 17](#_Toc191395976)

# **ВВЕДЕНИЕ**

Целью данной лабораторной работы является разработка асинхронного чат-бота для мессенджера Telegram, который использует языковые модели архитектур GPT и LLaMA для генерации ответов на вопросы пользователей. Чат-бот предназначен для помощи в учебных вопросах, таких как предоставление расписания, напоминаний о дедлайнах и ответов на образовательные запросы. В процессе выполнения работы были изучены и применены технологии асинхронного программирования, а также проведен сравнительный анализ языковых моделей GPT и LLaMA.

# **Задание на лабораторную работу**

## **1.1. Основная цель работы**

Основная цель работы — разработать асинхронный чат-бот для Telegram, который использует языковые модели GPT и LLaMA для генерации ответов на вопросы пользователей. Чат-бот должен предоставлять функциональность, связанную с учебным процессом, такую как показ расписания, напоминаний и ответов на образовательные вопросы.

## **1.2. Общие сведения**

Чат-боты — это программы, которые имитируют общение с пользователем и помогают решать различные задачи. В данной работе чат-бот разработан для помощи в учебном процессе. Основное отличие синхронного и асинхронного подхода заключается в том, что асинхронный подход позволяет обрабатывать несколько запросов одновременно, что повышает производительность и реактивность бота.

## **1.3. Языковые модели, рекомендуемые к использованию**

В работе использовались следующие языковые модели:

- LLaMA-2-7B: Модель с 7 миллиардами параметров, оптимизированная для чат-ботов.

- LLaMA-13B: Модель с 13 миллиардами параметров, обеспечивающая более высокую точность.

- LLaMA (RU): Модель с поддержкой русского языка.

- GPT2: Модель с низкими вычислительными требованиями.

- GPT-Neo: Открытая альтернатива GPT-3 с 125 миллионами параметров.

- ruGPT-3: Модель с поддержкой русского языка на основе GPT-2.

# **Выполнение лабораторной работы**

## **2.1. Выбор библиотеки для создания чат-бота**

Для разработки чат-бота была выбрана библиотека AsyncTeleBot, которая поддерживает асинхронное программирование и позволяет эффективно обрабатывать запросы пользователей.

## **2.2. Выбор предметной области**

Предметной областью чат-бота является учебный процесс. Чат-бот предоставляет функциональность для просмотра расписания, напоминаний о дедлайнах и ответов на образовательные вопросы.

## **2.3. Проектирование чат-бота**

Чат-бот имеет следующую функциональность:

- Показ расписания на текущий день.

- Показ списка дедлайнов.

- Ответы на образовательные вопросы с использованием языковых моделей GPT и LLaMA.

- Возможность выбора модели для генерации ответов.

## **2.4. Разработка асинхронного чат-бота**

Чат-бот разработан с использованием асинхронного подхода, что позволяет ему обрабатывать несколько запросов одновременно. Глубина вложенности функциональности составляет 5 уровней:

1. Главное меню.

2. Показ расписания.

3. Показ дедлайнов.

4. Выбор модели для генерации ответов.

5. Генерация ответов на образовательные вопросы.

## **2.5. Использованные технологии и библиотеки**

В процессе разработки чат-бота были использованы следующие технологии и библиотеки:

- AsyncTeleBot: Библиотека для создания асинхронного Telegram-бота.

- Hugging Face Transformers: Библиотека для работы с языковыми моделями GPT и LLaMA.

- Natasha: Библиотека для морфологического анализа русского языка.

- Dotenv: Для загрузки переменных окружения.

- Asyncio: Для реализации асинхронного программирования.

## **2.6. Сравнительный анализ языковых моделей**

В процессе работы были протестированы модели GPT и LLaMA. Основные выводы:

- LLaMA-2-7B, LLaMA-13B и LLaMA (RU) требуют Pro подписку, поэтому не удалось их протестировать.

- GPT2 и GPT-Neo показали хорошую производительность на простых запросах на английском, но для сложных задач их точность ниже, чем у моделей LLaMA.

- ruGPT-3 показал плохие результаты на русском языке и требует больше вычислительных ресурсов.

- LLaMa-3.2-3B-Instruct показал наилучший результат из всех доступных LLM.

## **2.7. Проблемы и ограничения**

Одной из основных проблем при использовании бесплатных моделей с Hugging Face является их низкая эффективность при обработке учебных запросов на русском языке. Только модель LLaMA Instruct показала удовлетворительные результаты. Для улучшения качества ответов требуется тонкая настройка (fine-tuning) моделей, что требует значительных вычислительных ресурсов, таких как GPU, которые недоступны на обычных ноутбуках.

## **2.8. Модулярность кода**

Код чат-бота был разработан с учетом принципов модулярности, что позволяет легко поддерживать и расширять функциональность бота. Каждый компонент системы выделен в отдельный модуль, что обеспечивает четкое разделение ответственности и упрощает тестирование и отладку. Основные модули, используемые в проекте:

1. handlers.py

Этот модуль содержит все обработчики сообщений и команд, которые взаимодействуют с пользователем. Каждая функция-обработчик отвечает за конкретную задачу, такую как показ расписания, напоминаний или генерация ответов на вопросы. Модуль также включает в себя логику для выбора модели и обработки учебных запросов.

2. models.py

В этом модуле реализована работа с языковыми моделями. Он отвечает за инициализацию моделей, их настройку и генерацию ответов на основе пользовательских запросов. Модуль использует библиотеку Hugging Face Transformers для взаимодействия с моделями GPT и LLaMA.

3. utils.py

Этот модуль содержит вспомогательные функции, такие как создание клавиатуры для меню, нормализация текста и проверка, является ли сообщение учебным. Также здесь реализована логика для загрузки переменных окружения и работы с морфологическим анализатором Natasha.

4. tbot.py

Основной модуль, который отвечает за запуск бота. Он инициализирует экземпляр бота, регистрирует обработчики из модуля `handlers.py` и запускает асинхронный цикл для обработки сообщений.

5. “.env”

Файл конфигурации, в котором хранятся чувствительные данные, такие как API-токены для Telegram и Hugging Face. Это позволяет избежать хранения таких данных в исходном коде и упрощает настройку бота для разных окружений.

**Преимущества модулярности:**

- Упрощение поддержки: Каждый модуль отвечает за свою часть функциональности, что упрощает внесение изменений и исправление ошибок.

- Повторное использование кода: Вспомогательные функции из `utils.py` могут быть легко использованы в других проектах.

- Тестируемость: Модули могут быть протестированы независимо друг от друга, что упрощает процесс тестирования.

- Масштабируемость: Добавление новой функциональности (например, новых команд или моделей) не требует изменения существующего кода, достаточно добавить новый модуль или расширить существующий.

## **2.9. Использование файла `educational\_keywords.txt`**

Для ограничения функциональности чат-бота только учебными темами был использован файл `educational\_keywords.txt`. Этот файл содержит список ключевых слов, связанных с учебным процессом, таких как "математика", "программирование", "лабораторная работа" и т.д. Когда пользователь отправляет сообщение, бот проверяет, содержит ли текст сообщения ключевые слова из этого списка. Если сообщение не содержит таких слов, бот отвечает, что может обрабатывать только учебные запросы.

**Как это работает:**

1. Загрузка ключевых слов:

При запуске бота ключевые слова загружаются из файла `educational\_keywords.txt` в память. Каждое слово приводится к нормализованной форме с использованием морфологического анализатора Natasha, чтобы учесть различные формы слов (например, "математика", "математике", "математику").

2. Проверка сообщений:

Когда пользователь отправляет сообщение, текст сообщения разбивается на слова, и каждое слово проверяется на соответствие ключевым словам из списка. Если хотя бы одно слово совпадает, сообщение считается учебным, и бот продолжает обработку запроса. В противном случае бот отвечает, что может обрабатывать только учебные вопросы.

**Преимущества:**

- Ограничение тематики: Бот фокусируется только на учебных вопросах, что помогает избежать обработки нерелевантных запросов.

- Гибкость: Список ключевых слов можно легко расширить или изменить, добавив новые слова в файл `educational\_keywords.txt`.

**Недостатки и костыльность решения:**

- Ограниченная точность: Это решение является костыльным, так как оно полагается на простой поиск ключевых слов. Оно не учитывает контекст сообщения, что может привести к ложным срабатываниям. Например, если пользователь напишет "Я не хочу учить математику", бот может посчитать это учебным запросом, хотя это не так.

- Зависимость от ключевых слов: Если пользователь использует синонимы или формулировки, которые не включены в список ключевых слов, бот может не распознать запрос как учебный.

- Отсутствие семантического анализа: Бот не понимает смысл сообщения, а только ищет совпадения по ключевым словам. Это ограничивает его способность корректно обрабатывать сложные запросы.

# **ВЫВОДЫ**

Чат-бот успешно справляется с поставленными задачами и может быть использован в учебном процессе для помощи студентам. Асинхронный подход позволяет эффективно обрабатывать запросы пользователей, а использование языковых моделей GPT и LLaMA обеспечивает высокое качество генерации ответов. Однако, для улучшения качества ответов на русском языке требуется тонкая настройка моделей, что требует значительных вычислительных ресурсов.

# **ИСПОЛЬЗУЕМАЯ ЛИТЕРАТУРА**

1. AsyncTeleBot Documentation – [https://github.com/eternnoir/pyTelegramBotAPI]

2. Hugging Face Transformers – [https://huggingface.co/docs/transformers/index]

3. Natasha: Russian NLP Library – [https://github.com/natasha/natasha]

4. Python Asyncio Documentation – [https://docs.python.org/3/library/asyncio.html]

5. PEP 8 — Style Guide for Python Code – [https://www.python.org/dev/peps/pep-0008/]

6. Hugging Face Model Hub – [https://huggingface.co/models]

7. https://school.kontur.ru/publications/2567

# **ПРИЛОЖЕНИЕ А. Работа чат-бота**

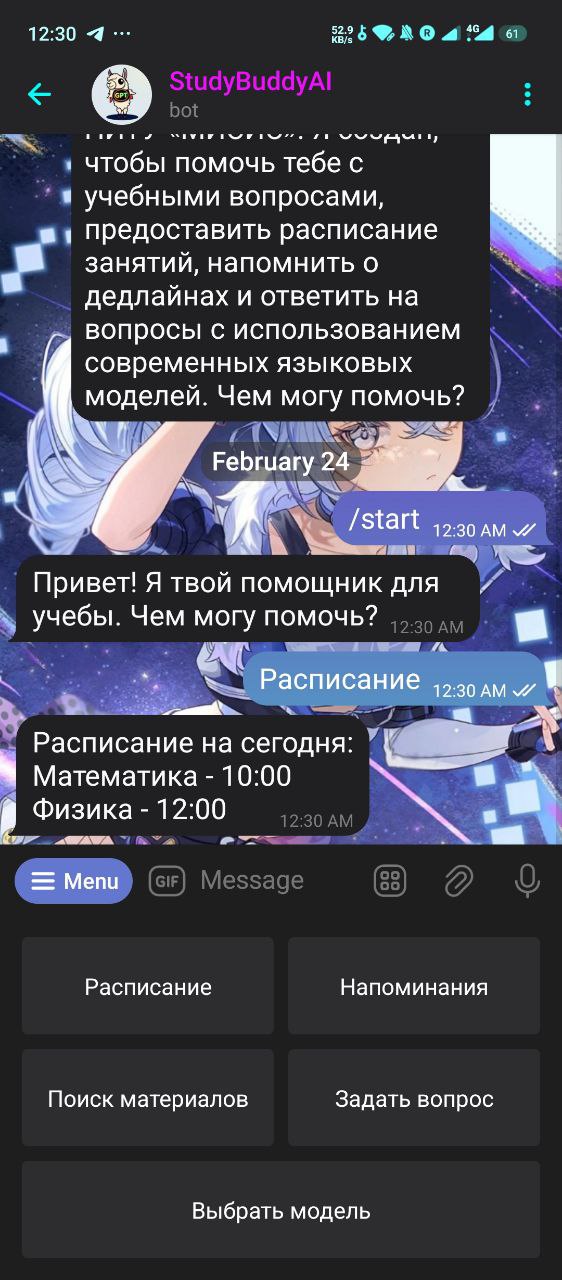
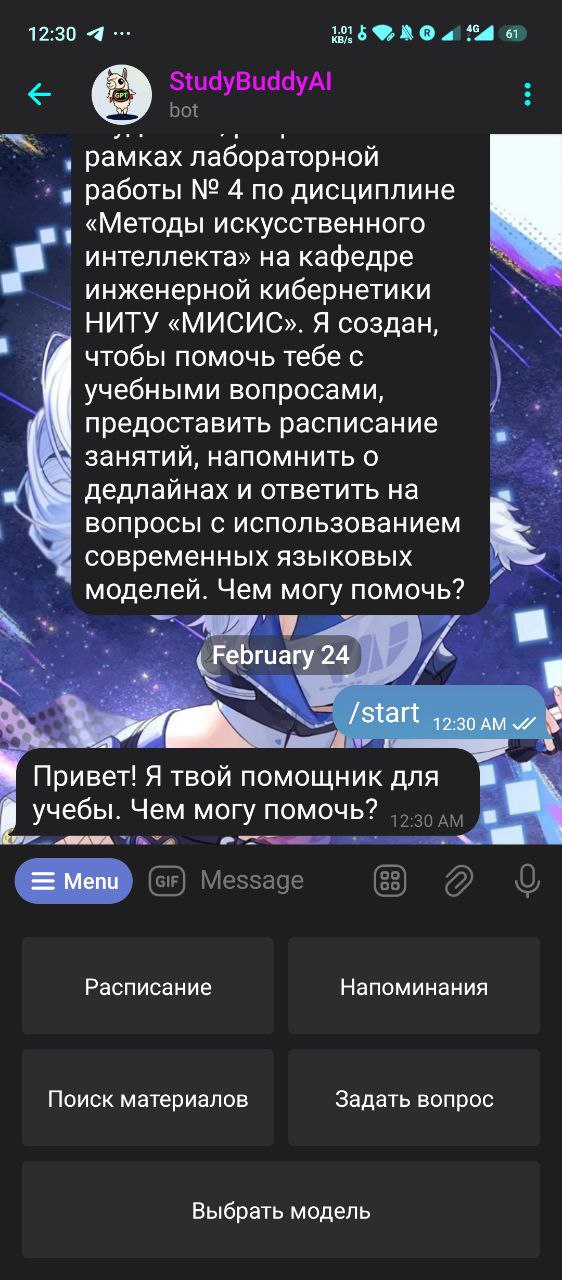
******Скриншоты работы чат-бота:**

Рисунок 1. Главное меню чат-бота

Рисунок 2. Показ расписания

Примечание: все функции кроме выбора модели LLM и генерации ответа условны, могут быть реализованы по-настоящему в будущем.

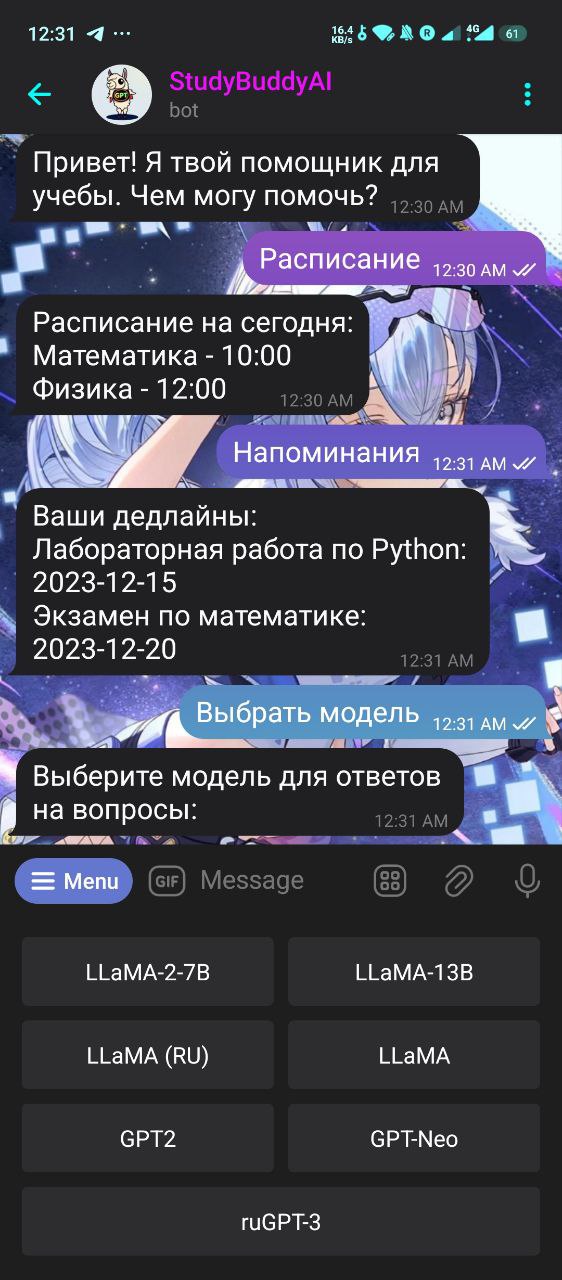
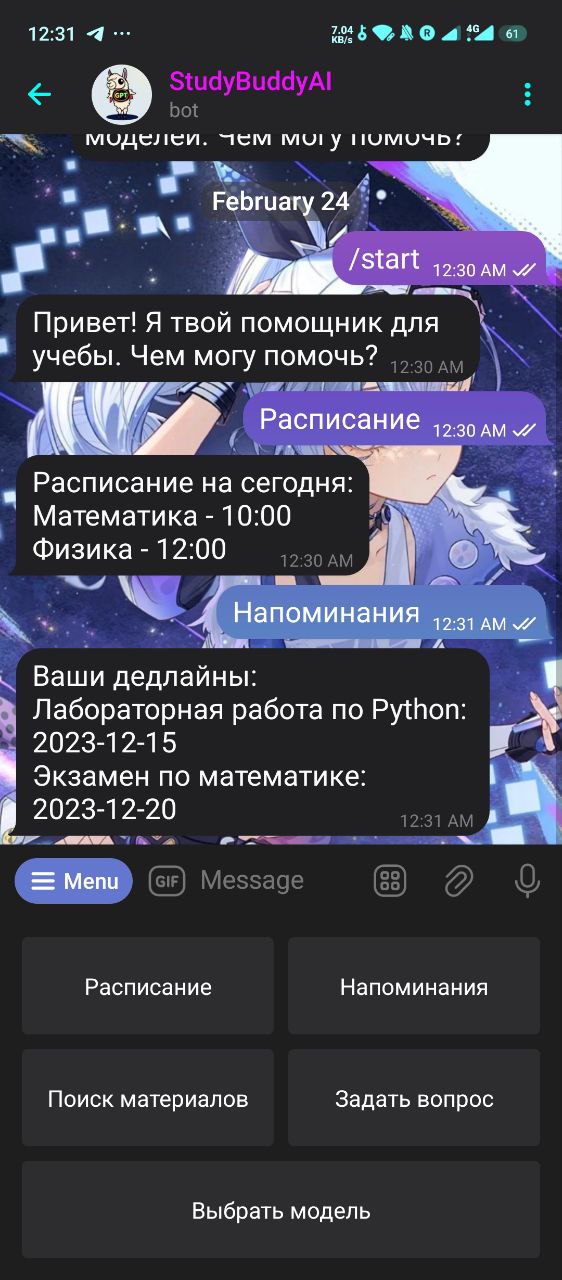
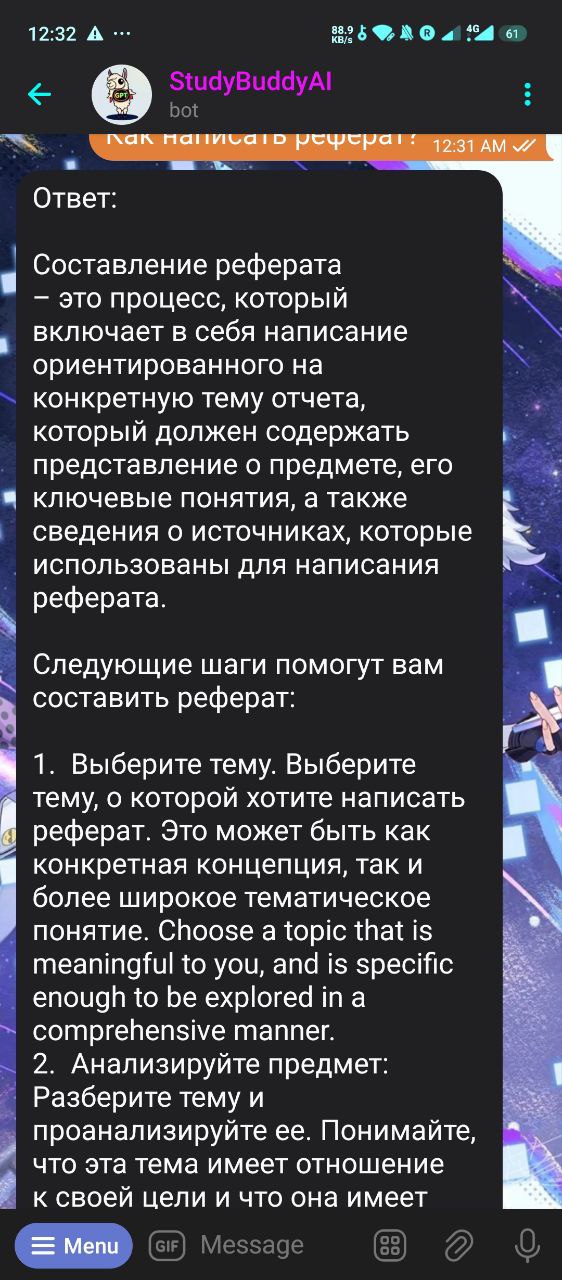


Рисунок 4. Выбор модели

Рисунок 3. Показ дедлайнов



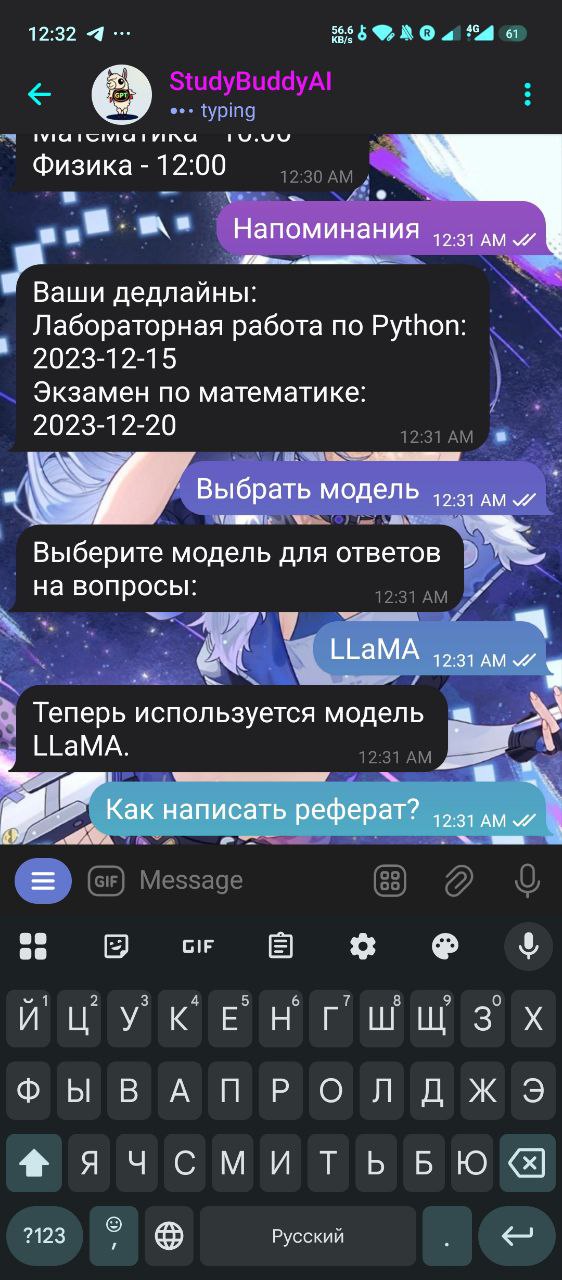


Рисунок 6. Ответ

Рисунок 5. Генерация ответа

# **ПРИЛОЖЕНИЕ Б. Код чат-бота**

**tbot.py**

import asyncio

from telebot.async\_telebot import AsyncTeleBot

from handlers import register\_handlers

from utils import load\_env\_vars

# Загрузка переменных окружения

API\_TOKEN = load\_env\_vars("TELEGRAM\_BOT\_TOKEN")

# Инициализация бота

bot = AsyncTeleBot(API\_TOKEN)

# Регистрация обработчиков

register\_handlers(bot)

async def main() -> None:

    """Запуск бота."""

    await bot.polling()

if \_\_name\_\_ == "\_\_main\_\_":

    asyncio.run(main())

**utils.py**

import os

from typing import Optional, List

from dotenv import load\_dotenv

from telebot.types import ReplyKeyboardMarkup, KeyboardButton

from natasha import MorphVocab

import unicodedata

import logging

logging.basicConfig(level=logging.DEBUG)

# Инициализация морфологического словаря

morph\_vocab = MorphVocab()

def load\_env\_vars(key: str) -> str:

    """

    Загружает переменную окружения по ключу.

    Аргументы:

        key (str): Ключ переменной окружения.

    Возвращает:

        str: Значение переменной окружения.

    Исключения:

        ValueError: Если переменная окружения не найдена.

    """

    load\_dotenv(".env")

    value = os.getenv(key)

    if value is None:

        raise ValueError(f"Переменная окружения {key} не найдена.")

    return value

def create\_menu\_keyboard() -> ReplyKeyboardMarkup:

    """

    Создает клавиатуру с основными командами для пользователя.

    Возвращает:

        ReplyKeyboardMarkup: Объект клавиатуры с кнопками.

    """

    markup = ReplyKeyboardMarkup(row\_width=2)

    markup.add(

        KeyboardButton("Расписание"),

        KeyboardButton("Напоминания"),

        KeyboardButton("Поиск материалов"),

        KeyboardButton("Задать вопрос"),

        KeyboardButton("Выбрать модель"),

    )

    return markup

def normalize\_text(text: str) -> str:

    """

    Нормализовать текст, чтобы обеспечить согласованное сравнение.

    """

    return unicodedata.normalize('NFKC', text)

def is\_educational\_message(message\_text: str) -> bool:

    """

    Проверяет, связано ли сообщение с учебой, используя список ключевых слов и Natasha для учета склонений.

    """

    # Чтение ключевых слов из файла

    try:

        with open("educational\_keywords.txt", "r", encoding="utf-8") as file:

            educational\_keywords = [

                normalize\_text(line.strip().lower()) for line in file.readlines() if line.strip()

            ]

    except FileNotFoundError:

        logging.error("Файл educational\_keywords.txt не найден.")

        return False

    if not educational\_keywords:

        logging.error("Файл с ключевыми словами пуст.")

        return False

    # Приведение текста сообщения к нижнему регистру и нормализация

    message\_text = normalize\_text(message\_text.lower())

    # Разбиваем сообщение на слова

    user\_words = message\_text.split()

    for word in user\_words:

        # Приводим слово к начальной форме (лемматизация)

        parsed\_word = morph\_vocab(word)  # Вернуть список объектов MorphForm

        if parsed\_word:  # Убедиться что список не пуст

            # Вытаскиваем нормальные формы слова как string

            normal\_forms = [form.normal\_form for form in parsed\_word if hasattr(form, 'normal\_form')]

            # Проверить все нормальные формы слова

            for normal\_form in normal\_forms:

                normal\_form = normalize\_text(normal\_form)

                # Логирование normal\_form для дебага

                logging.debug(f"Word: {word}, Normal Form: {normal\_form}")

                # Проверяем, есть ли начальная форма в списке ключевых слов

                if normal\_form in educational\_keywords:

                    return True

    return False

**handlers.py**

from typing import Dict, List

from telebot import types

from telebot.async\_telebot import AsyncTeleBot

from models import set\_model, generate\_response, current\_model

from utils import create\_menu\_keyboard, is\_educational\_message

# Пример расписания

schedule: Dict[str, List[str]] = {

    "Понедельник": ["Математика - 10:00", "Физика - 12:00"],

    "Вторник": ["Программирование - 9:00", "Английский - 14:00"],

}

# Пример дедлайнов

deadlines: Dict[str, str] = {

    "Лабораторная работа по Python": "2023-12-15",

    "Экзамен по математике": "2023-12-20",

}

# Словарь для сопоставления названий кнопок с именами моделей

model\_mapping: Dict[str, str] = {

    "LLaMA-2-7B": "meta-llama/Llama-2-7b-chat-hf",

    "LLaMA-13B": "meta-llama/Llama-2-13b-chat-hf",

    "LLaMA (RU)": "IlyaGusev/saiga\_llama3\_8b",

    "LLaMA": "meta-llama/Llama-3.2-3B-Instruct",

    "GPT2": "openai-community/gpt2",

    "GPT-Neo": "EleutherAI/gpt-neo-125m",

    "ruGPT-3": "ai-forever/rugpt3medium\_based\_on\_gpt2",

}

def register\_handlers(bot: AsyncTeleBot) -> None:

    """

    Регистрирует обработчики команд и сообщений.

    Аргументы:

        bot (AsyncTeleBot): Экземпляр бота.

    """

    @bot.message\_handler(commands=['start'])

    async def send\_welcome(message: types.Message) -> None:

        """

        Отправляет приветственное сообщение и показывает меню.

        Аргументы:

            message (types.Message): Объект сообщения от пользователя.

        """

        await bot.send\_message(

            message.chat.id,

            "Привет! Я твой помощник для учебы. Чем могу помочь?",

            reply\_markup=create\_menu\_keyboard(),

        )

    @bot.message\_handler(func=lambda message: message.text == "Расписание")

    async def show\_schedule(message: types.Message) -> None:

        """

        Показывает расписание на сегодня.

        Аргументы:

            message (types.Message): Объект сообщения от пользователя.

        """

        today = "Понедельник"  # Замените на реальное определение дня

        if today in schedule:

            await bot.send\_message(

                message.chat.id,

                f"Расписание на сегодня:\n" + "\n".join(schedule[today]),

            )

        else:

            await bot.send\_message(message.chat.id, "Сегодня занятий нет.")

    @bot.message\_handler(func=lambda message: message.text == "Напоминания")

    async def show\_deadlines(message: types.Message) -> None:

        """

        Показывает список дедлайнов.

        Аргументы:

            message (types.Message): Объект сообщения от пользователя.

        """

        reminders = [f"{task}: {date}" for task, date in deadlines.items()]

        await bot.send\_message(

            message.chat.id,

            "Ваши дедлайны:\n" + "\n".join(reminders),

        )

    @bot.message\_handler(func=lambda message: message.text == "Выбрать модель")

    async def choose\_model(message: types.Message) -> None:

        """

        Показывает клавиатуру для выбора модели.

        Аргументы:

            message (types.Message): Объект сообщения от пользователя.

        """

        markup = types.ReplyKeyboardMarkup(row\_width=2)

        markup.add(\*[types.KeyboardButton(model\_name) for model\_name in model\_mapping.keys()])

        await bot.send\_message(

            message.chat.id,

            "Выберите модель для ответов на вопросы:",

            reply\_markup=markup,

        )

    @bot.message\_handler(func=lambda message: message.text in model\_mapping.keys())

    async def handle\_model\_selection(message: types.Message) -> None:

        """

        Устанавливает выбранную модель.

        Аргументы:

            message (types.Message): Объект сообщения от пользователя.

        """

        selected\_model\_name = model\_mapping.get(message.text, None)

        if selected\_model\_name:

            set\_model(selected\_model\_name)

            await bot.send\_message(

                message.chat.id,

                f"Теперь используется модель {message.text}.",

                reply\_markup=create\_menu\_keyboard(),

            )

        else:

            await bot.send\_message(

                message.chat.id,

                "Неизвестная модель. Пожалуйста, выберите модель из списка.",

                reply\_markup=create\_menu\_keyboard(),

            )

    @bot.message\_handler(func=lambda message: True)

    async def handle\_message(message: types.Message) -> None:

        """

        Обрабатывает текстовые сообщения пользователя.

        Аргументы:

            message (types.Message): Объект сообщения от пользователя.

        """

        user\_input = message.text

        # Проверяет, связано ли сообщение с учебой

        if is\_educational\_message(user\_input):

            try:

                 # Отображаем статус "Печатает..."

                await bot.send\_chat\_action(message.chat.id, "typing")

                #await asyncio.sleep(1)  # Симулируем задержку в 1 секунду

                response = generate\_response(user\_input)

                await bot.send\_message(message.chat.id, f"Ответ:\n{response}")

            except Exception as e:

                await bot.send\_message(message.chat.id, f"Ошибка при запросе к модели: {str(e)}")

        else:

            await bot.send\_message(message.chat.id, "Пожалуйста, задавайте вопросы, связанные с учебой.")

**models.py**

from huggingface\_hub import InferenceClient

from utils import load\_env\_vars

# Загрузка переменных окружения

HUGGING\_FACE\_API\_TOKEN = load\_env\_vars("HUGGING\_FACE\_API\_TOKEN")

# Инициализация Hugging Face Inference Client

client = InferenceClient(api\_key=HUGGING\_FACE\_API\_TOKEN)

# Переменная для хранения выбранной модели

current\_model: str = "meta-llama/Llama-3.2-3B-Instruct"  # Llama-3.2-3B-Instruct

def set\_model(model\_name: str) -> None:

    """

    Устанавливает текущую модель для генерации ответов.

    Аргументы:

        model\_name (str): Название модели.

    """

    global current\_model

    current\_model = model\_name

def generate\_response(prompt: str) -> str:

    """

    Генерирует ответ на запрос пользователя с использованием текущей модели.

    Аргументы:

        prompt (str): Запрос пользователя.

    Возвращает:

        str: Ответ модели.

    """

    response = client.text\_generation(

        prompt=prompt,

        model=current\_model,

        max\_new\_tokens=700,

        temperature=0.7,

        top\_k=50,

        do\_sample=True,

    )

    return response