### **Введение**

В последние годы наблюдается стремительный рост интереса к технологиям искусственного интеллекта и их применению для решения прикладных задач. Одним из наиболее динамично развивающихся направлений является создание интеллектуальных диалоговых систем, или чат-ботов, способных взаимодействовать с пользователем на естественном языке. Современные чат-боты эволюционировали от простых систем, работающих по жестко заданным сценариям, до сложных ассистентов, использующих модели обработки естественного языка (NLP) для понимания намерений пользователя и предоставления релевантной информации.

Особый интерес представляет применение таких систем в узкоспециализированных областях, таких как туризм, экология и образование. Предоставление точной, верифицированной и, что немаловажно, наглядной информации является ключевой задачей для повышения осведомленности и интереса к уникальным природным объектам. Байкальский регион, с его эндемичной флорой и фауной, представляет собой идеальную предметную область для применения подобных технологий. Традиционные источники информации, такие как справочники или веб-сайты, часто не обладают необходимой интерактивностью и не способны отвечать на комплексные, контекстуально-зависимые вопросы пользователей.

Разработка специализированного чат-бота, ориентированного на экосистему Байкала, позволяет решить эту проблему, предоставляя пользователю мощный инструмент для исследования региона. Однако создание такого ассистента сопряжено с рядом технических вызовов. Необходимо не только обеспечить точное распознавание запросов, но и реализовать эффективное взаимодействие с базами данных, содержащими разнородную информацию: текстовые описания, изображения и геопространственные данные. Современный подход к решению таких задач заключается в использовании гибридных архитектур, сочетающих сильные стороны классических NLU-фреймворков и больших языковых моделей (LLM).

**Цель проекта:**

Разработать Telegram-бота "Эко-ассистент Байкала", способного предоставлять пользователям мультимедийную информацию (текст, изображения, интерактивные карты) о флоре и фауне Байкальского региона, основываясь на анализе запросов на естественном языке.

**Задачи для достижения цели (Пока тут бред!):**

1. **Исследовать** существующие подходы к созданию интеллектуальных диалоговых систем, включая классические NLU-фреймворки и системы на основе LLM.
2. **Проанализировать** архитектурные решения для построения гибридных чат-ботов и выбрать оптимальный стек технологий.
3. **Спроектировать** микросервисную архитектуру приложения, включающую компоненты для взаимодействия с Telegram, обработки NLU, доступа к базам данных и интеграции с LLM.
4. **Реализовать** два независимых пайплайна обработки запросов: один на базе Rasa для структурированных команд и второй на базе GigaChat для анализа запросов на естественном языке.
5. **Разработать** механизм фоллбэка для повышения полноты ответов в случаях отсутствия информации в основной базе знаний.
6. **Протестировать** и оценить работоспособность и производительность реализованной системы.

### **1. Анализ**

#### **1.1 Диалоговые системы и их эволюция**

Диалоговые системы, или чат-боты, представляют собой программные комплексы, предназначенные для имитации осмысленного разговора с пользователем. Исторически первые чат-боты, такие как ELIZA (1966), работали на основе простого сопоставления с шаблонами и не обладали реальным "пониманием" языка. С развитием технологий обработки естественного языка (Natural Language Processing, NLP) чат-боты стали значительно сложнее.

Современные системы можно условно разделить на две большие категории:

* **Декларативные (Rule-Based):** Работают по заранее определенным правилам и сценариям. Они предсказуемы, надежны и быстры, но их возможности строго ограничены заложенными в них скриптами. Такие боты часто используются в службах поддержки для ответов на типовые вопросы.
* **Интеллектуальные (AI-Based):** Используют машинное обучение и NLP для анализа и понимания запросов пользователя. Они способны распознавать намерения (intents), извлекать сущности (entities) и поддерживать более гибкий диалог.

#### **1.2 Архитектура современных NLU-ассистентов**

Ключевым компонентом интеллектуального чат-бота является модуль NLU (Natural Language Understanding). Его основная задача — преобразовать неструктурированный текст пользователя в структурированный формат, понятный машине. Стандартный NLU-пайплайн включает в себя несколько этапов:

1. **Токенизация:** Разбиение текста на отдельные слова или символы (токены).
2. **Извлечение сущностей (Entity Extraction):** Распознавание в тексте важных фрагментов, таких как имена, даты, географические названия (например, "кедр", "зимой", "Листвянка").
3. **Определение намерения (Intent Classification):** Классификация всего запроса пользователя для определения его основной цели (например, получить картинку, получить информацию).

Популярным фреймворком для построения таких систем является **Rasa Open Source**. Он предоставляет инструменты для создания NLU-моделей и управления диалогом (Dialogue Management), позволяя разработчику полностью контролировать логику бота.

#### **1.3 Большие языковые модели (LLM) в диалоговых системах**

С появлением больших языковых моделей, таких как GPT-4 или GigaChat, произошел качественный скачок в возможностях диалоговых систем. В отличие от классических NLU-моделей, которые обучаются на конкретных примерах для распознавания ограниченного набора интентов и сущностей, LLM обладают "общим" пониманием языка.

В контексте чат-ботов LLM могут использоваться для решения двух основных задач:

1. **Анализ и структурирование запроса:** LLM может выступить в роли "универсального NLU-модуля". Ему можно дать на вход сырой текст пользователя и попросить вернуть структурированный JSON-объект с намерением и извлеченными параметрами. Этот подход обеспечивает невероятную гибкость, так как система способна понимать запросы, которые не были явно предусмотрены в обучающих данных.
2. **Генерация ответа:** LLM может генерировать человекоподобные текстовые ответы, что особенно полезно для фоллбэк-сценариев, когда в основной базе знаний нет готового ответа.

### **1.4 Архитектура с переключаемыми пайплайнами обработки**

Для решения поставленных задач в рамках проекта "Эко-ассистент Байкала" была спроектирована и реализована архитектура, основанная на **двух независимых, переключаемых пайплайнах (режимах) обработки запросов**. Такой подход позволяет пользователю самому выбирать способ взаимодействия с системой в зависимости от его потребностей, сочетая сильные стороны различных технологий.

Система включает в себя следующие режимы:

1. **Режим Rasa:** В данном режиме вся обработка запроса делегируется классическому NLU-фреймворку Rasa Open Source. Этот пайплайн оптимизирован для распознавания заранее определенных, структурированных команд. Пользовательский ввод проходит через NLU-модель, которая извлекает намерение (intent) и сущности (entities), после чего система выполняет соответствующее действие (action). Этот режим обеспечивает высокую скорость и предсказуемость для стандартных и частотных запросов.
2. **Режим GigaChat:** В этом режиме система использует большую языковую модель (LLM) GigaChat для семантического анализа запроса. Вместо того чтобы сопоставлять запрос с заранее определенными шаблонами, основной бот отправляет сырой текст пользователя в LLM со специальным системным промптом. Задача LLM — не сгенерировать ответ, а преобразовать неструктурированный запрос в структурированную JSON-команду, которую затем исполняет бот. Этот подход обеспечивает высокую гибкость и позволяет понимать сложные, вариативные формулировки на естественном языке, которые не были предусмотрены в обучающих данных.

Пользователь может в любой момент переключиться между этими двумя режимами через меню настроек бота. Несмотря на различие в логике обработки, оба пайплайна обращаются к единому бэкенд-сервису для получения фактических данных (текстов, изображений, карт) и могут использовать общий микросервис для фоллбэк-сценариев. Такая архитектура позволяет экспериментально сравнивать эффективность двух различных подходов к созданию диалоговых систем в рамках одного приложения.

### **1.5 Проблема и актуальность разработки**

**Проблема:** Существующие источники информации о флоре и фауне Байкальского региона (веб-сайты, справочники, общие ассистенты) обладают рядом фундаментальных недостатков. Они либо неинтерактивны и не могут отвечать на комплексные, уточняющие вопросы (как, например, печатные издания или статические сайты), либо не обладают специализированной и верифицированной базой данных, предоставляя общую или неточную информацию (как общие ассистенты, склонные к "галлюцинациям"). Отсутствует единый инструмент, который бы сочетал **глубину специализированных знаний** с **гибкостью современного диалогового интерфейса**, ориентированного на мультимедийный контент.

**Актуальность:** Разработка "Эко-ассистента Байкала" является актуальной, поскольку напрямую решает обозначенную проблему. Проект предлагает не один, а **два альтернативных подхода к пониманию естественного языка** в рамках единого приложения, позволяя пользователю выбрать наиболее удобный для него:

1. **Классификационный режим (Rasa):** Использует NLU-модель, специально обученную на наборе примеров для точной классификации заранее определенных типов запросов (намерений) и извлечения из них параметров. Сильной стороной этого подхода является высокая надежность и предсказуемость при выполнении типовых, заранее спроектированных задач.
2. **Семантический режим (GigaChat):** Использует большую языковую модель для динамической интерпретации смысла произвольных запросов пользователя. Этот подход не ограничен набором обучающих примеров и позволяет понимать более сложные и вариативные формулировки, семантически анализируя текст "на лету".

Создание такой системы с переключаемыми режимами обработки позволяет не только предоставить пользователям уникальный инструмент для исследования экосистемы Байкала, но и провести практическое сравнение эффективности двух ведущих подходов к построению диалоговых систем. Актуальность проекта подкрепляется его высоким потенциалом в сферах туризма, образования и эко-просвещения.

### **Обзор существующих программных средств**

Для определения уникальности и актуальности разрабатываемого "Эко-ассистента Байкала" необходимо провести анализ существующих на рынке программных решений, которые частично или полностью пересекаются с его функциональностью. Анализ будет проведен по трем ключевым категориям: общие интеллектуальные ассистенты, специализированные приложения-определители и традиционные информационные чат-боты. Оценка будет проводиться по адаптированным критериям, релевантным для диалоговых систем.

#### **2.1 Общие интеллектуальные ассистенты (Яндекс Алиса, Google Assistant)**

Данная категория представляет собой наиболее технологически продвинутые диалоговые системы, способные поддерживать разговор на широкий круг тем.

* **2.1.1 Общая оценка интерфейса.** Интерфейс является преимущественно голосовым, но также поддерживает текстовый ввод. Взаимодействие максимально приближено к естественному человеческому диалогу. Системы отлично справляются с поддержанием контекста и обработкой сложных, многосоставных предложений.
* **2.1.2 Объем и структура представленной информации.** Объем информации практически неограничен, так как ассистенты используют для ответов всю проиндексированную сеть Интернет. Однако это является и их недостатком: информация не является специализированной и часто представляет собой краткую выдержку из первого найденного источника (например, Википедии). Структура ответа, как правило, — это короткий текстовый блок, иногда сопровождаемый ссылкой или изображением из поиска. Мультимедийные возможности ограничены показом статичных картинок и не включают генерацию интерактивных карт по запросу.
* **2.1.3 Наличие и структура меню.** Меню как таковое в классическом понимании отсутствует. Навигация осуществляется полностью через языковые команды. Существуют стандартные команды для вызова справки, но нет структурированных меню для навигации по узкоспециализированной предметной области.
* **2.1.4 Удобство форм для ввода информации.** Основной формой ввода является текстовая строка или голосовой запрос. Благодаря использованию мощных LLM, эти системы обладают высочайшей гибкостью в понимании естественного языка, синонимов и различных формулировок одного и того же вопроса.
* **2.1.5 Возможность поиска информации.** Поиск является основной функцией, но он носит общий характер. Ассистенты не подключены к специализированным, верифицированным базам данных, что в контексте научной информации о флоре и фауне может приводить к предоставлению неточной, устаревшей или откровенно ложной информации ("галлюцинациям").
* **2.1.6 Общий вывод по категории.** Общие ассистенты предоставляют лучший на рынке опыт естественного диалога, но не могут служить надежным источником для получения специализированных знаний. Их функциональность не приспособлена для решения узких задач, таких как построение карт ареалов или поиск изображений по сложным признакам.

#### **2.2 Специализированные приложения-определители (Picture This, iNaturalist)**

Это мобильные приложения, основная задача которых — идентификация видов растений и животных по фотографии пользователя.

* **2.2.1 Общая оценка интерфейса.** Интерфейс является графическим (GUI), а не диалоговым. Взаимодействие с пользователем происходит через элементы управления: кнопки, вкладки, экраны. Интерфейс интуитивно понятен для своей основной задачи — загрузки и анализа фотографий.
* **2.2.2 Объем и структура представленной информации.** Эти приложения обладают огромными, хорошо структурированными и верифицированными базами данных. Информация по каждому виду включает таксономию, качественные фотографии, подробные описания и, в случае iNaturalist, карту с точками наблюдений от других пользователей. Данные представлены в формате карточек, что очень удобно для восприятия.
* **2.2.3 Наличие и структура меню.** Навигация осуществляется через стандартные для мобильных приложений меню (например, нижняя панель вкладок), которые позволяют переключаться между функциями: идентификация, личная коллекция, карта и т.д.
* **2.2.4 Удобство форм для ввода информации.** Основной "формой ввода" является камера или галерея устройства. Текстовый ввод используется только для поиска по названию вида и не предполагает обработки запросов на естественном языке. Задать вопрос вроде "какие хвойные деревья растут на берегу Байкала?" невозможно.
* **2.2.5 Возможность поиска информации.** Поиск ограничен либо названием вида, либо идентификацией по фото. Сложные, многокритериальные запросы не поддерживаются.
* **2.2.6 Общий вывод по категории.** Приложения-определители являются превосходным источником достоверной и хорошо структурированной информации. Однако они не являются диалоговыми системами и не способны обеспечить интерактивное исследование предметной области через вопросы на естественном языке.

#### **2.3 Традиционные информационные чат-боты (банковские, справочные)**

К этой категории относятся чат-боты, работающие по заранее заданным сценариям и правилам, часто встречающиеся на сайтах компаний для поддержки клиентов.

* **2.3.1 Общая оценка интерфейса.** Интерфейс строго функционален, часто основан на кнопочных меню. Диалог ощущается как "механический" и негибкий.
* **2.3.2 Объем и структура представленной информации.** Объем информации строго ограничен базой знаний, заложенной разработчиками. Ответы представляют собой заранее написанные текстовые блоки.
* **2.3.3 Наличие и структура меню.** Меню является основным элементом навигации. Оно имеет древовидную структуру и позволяет пользователю предсказуемо перемещаться по доступным опциям.
* **2.3.4 Удобство форм для ввода информации.** Понимание естественного языка у таких ботов крайне ограничено или отсутствует. Они хорошо распознают ключевые слова ("доставка", "цена"), но любой отход от ожидаемого формата запроса приводит к ошибке "я вас не понял".
* **2.3.5 Возможность поиска информации.** Поиск основан на простом сопоставлении с ключевыми словами или навигации по базе знаний через меню.
* **2.3.6 Общий вывод по категории.** Традиционные чат-боты надежны и быстры в рамках своих узких, заранее определенных задач. Однако они абсолютно не подходят для роли исследовательского инструмента, так как не обладают гибкостью для понимания разнообразных вопросов пользователя.

#### **Общий вывод**

Проведенный анализ существующих программных средств показывает, что на рынке отсутствует решение, которое бы объединяло сильные стороны всех рассмотренных категорий. Общие ассистенты предлагают гибкий диалог, но не имеют доступа к специализированной и достоверной базе данных. Приложения-определители обладают такой базой, но лишены диалогового интерфейса. Традиционные чат-боты слишком ограничены в своих возможностях. Ценность и новизна "Эко-ассистента Байкала" заключается в том, что он спроектирован для заполнения именно этой ниши. Он сочетает гибкость понимания естественного языка, характерную для современных ассистентов, со специализированной, верифицированной базой знаний, дополняя это уникальными мультимедийными возможностями, такими как генерация интерактивных карт по запросу. Это делает его не просто очередным чат-ботом, а полноценным интерактивным инструментом для исследования.