Использование ChatGPT API в приложениях

Понять как работают Large language Models (LLM)

# Приветствие

Всем привет.

Меня зовут Константин Волошенко, я бизнес - аналитик из Санкт-Петербурга.

Одно из моих увлечений — нейронные сети и искусственный интелект.

Сегодня я расскажу о том, как вы можете интегрировать ChatGPT в свое приложение.

Вступление

# Я представлю обзор того, что такое ChatGPT и как она работает, а затем объясню, как вы можете использовать API ChatGPT для включения его в свое программное обеспечение.

# Что такое и как работает ChatGPT

ChatGPT — это языковая модель, разработанная OpenAI. Она использует метод, называемый Generative Pre-trained Transformer (GPT) для генерации текстовых ответов, которые напоминают человеческую речь. Эта модель была обучена на большом количестве текста из Интернета, что позволило ей изучить шаблоны, грамматику и даже некоторое понимание различных тем. Она работает в диалоговом режиме, когда пользователи могут вводить вопросы, а ChatGPT генерирует соответствующий ответ.

# Шаг 1. Простой запрос. Использование ChatGPT API в приложениях

В первую очередь, чтобы интегрироваться с языковой моделью, мы должны найти подходящий API. Да, есть такой API для ChatGPT.

Как взаимодействовать с ChatGPT через API? ChatGPT использует концепцию "ролей" для различных участников в диалоге. Каждое сообщение в списке messages является словарем, имеющим два поля: role и content. Role может принимать одно из трех значений: "system", "user" или "assistant", указывая, кто именно "говорит" сообщение. Content содержит само сообщение.

В стандартном случае, три основные роли могут быть включены в чат:

* "**system**": Системная роль используется для установки контекста диалога. Это как общая инструкция для модели, которая определяет её поведение во время диалога. Сообщения от "system" обычно помещаются в начале диалога и могут содержать информацию, как модель должна вести себя в роли ассистента.
* "**user**": Это роль для человека- пользователя, который ведет диалог с моделью. Ваши запросы к модели будут выставлены под этой ролью.
* "**assistant**": Это роль для модели AI, которая отвечает на запросы от пользователя.

См. более подробно здесь:

<https://platform.openai.com/docs/guides/gpt/chat-completions-ap>

# Шаг 2. Промт и База Знаний

Во-вторых, мы должны решить, как именно отправлять наши фактические данные в языковую модель. Итак, мы будем использовать обученную языковую модель в качестве преобразователя, преобразуя наши данные в необходимую форму ответа.

Для этого заранее подготавливаем файлы с промтом и базой знаний.

Чтобы модель смогла адекватно отвечать на вопросы клиентов о вашем предприятии, необходимо подготовить Промт и Базу знаний.

## Промт

При создании промптов (инструкций) для модели ChatGPT есть несколько принципов, которые могут помочь вам получить лучшие результаты:

* **Конкретизация**: Чем конкретнее ваш промпт, тем более конкретный ответ вы получите. Ваши инструкции должны быть понятными и конкретными. Если инструкции слишком общие или неоднозначные, модель может не понять, что вы от нее хотите.
* Роль **system**: Использование роли **system** для указания поведения модели может быть очень полезным. Например, вы можете указать в system: "Вы - помощник, который специализируется на истории искусства и всегда стремится предоставить подробные и точные ответы". Это сообщение задает контекст для модели. Роль **system** позволяет установить задачу или дать инструкции модели. Это сообщение видит модель, но оно не считается частью диалога с пользователем.

## База знаний

Структура вашей Базы Знаний играет жизненно важную роль в предоставлении ответов на запросы пользователей. Основной целью структуры базы знаний является обеспечение ее эффективного функционирования.

Вот несколько ключевых принципов, которые следует учесть при формировании структуры базы знаний:

* **Категоризация и классификация**: База знаний должна быть разделена на ясные и логически организованные блоки, соответствующие различным областям знаний или типам вопросов клиентов. Например, это могут быть разделы, связанные с продуктами, услугами, поддержкой, вопросами оплаты и т.д. В каждом блоке могут быть подблоки для дополнительной детализации. Это позволит LangChain легко найти релевантные чанки, соответствующие запросу.
* **Иерархия**: База знаний может быть организована в виде иерархической структуры, где более общие темы располагаются на верхнем уровне, а более специфические вопросы и ответы находятся на более низких уровнях. Так проще контролировать наполненность базы знаний нужной информацией и обновлять ее при необходимости.

С помощью этих принципов мы можем гарантировать, что база знаний организована, логична и совместима с моделями.

Однако мы не можем передать достаточно большую базу знаний в языковую модель и эта модель не доступна нам для до-обучения.

# Шаг 3: Телеграм бот. Собираем все вместе

В-третьих, нам нужно интегрировать взаимодействие с языковой моделью в наше приложение.

Данная диаграмма описывает процесс интеграции Телеграм-бота и ChatGPT для предоставления ответов на вопросы пользователей.

1. Приложение загружает файл с Промтом и файл с Базой Знаний из Google Drive.
2. Приложение разбивает Базу Знаний на чанки с помощью CharacterTextSplitter.
3. Приложение создает индексную базу данных (Индексная DB) на основе чанков Базы Знаний.
4. Индексная база данных преобразует чанки Базы Знаний в эмбеддинги для дальнейшей обработки.
5. Далее асинхронно:.
   1. Пользователь отправляет вопрос Telegram Bot.
   2. Приложение получает вопрос пользователя.
   3. Приложение ищет соответствующие фрагменты в базе знаний на основе вопроса пользователя.
   4. Приложение отправляет запрос в ChatGPT, включая промт, содержимое сообщения (вопрос и соответствующие фрагменты) и параметр температуры, управляющий случайностью ответа.
   5. ChatGPT генерирует ответ и отправляет его обратно в приложение.
   6. В итоге ответ от ChatGPT отправляется обратно пользователю в Телеграмм.

## Примеры вопросов

|  |  |
| --- | --- |
| Вопрос | Категория |
| ты кто? | Общий |
| цены? | База знаний |
| Есть ли у вас скидки? | База знаний |
| скидка студентам? | База знаний |
| Что про алкоголь? | База знаний |
| курить можно? | База знаний |
| вы все козлы | Стресс устойчивость |
| А как вас найти? | База знаний |
| Какие у вас есть игры? | База знаний |
| What games do you have? | Другой язык |
| ¿Qué juegos tienes? | Другой язык |

# Выводы

1. Для интеграции с языковой моделью нужно использовать API.
2. Нужно составить Промт и Базу знаний.
3. Телеграмм бот - очень простой способ интеграции.
4. Кроме ChatGPT есть и другие модели, например семейство моделей Llama (Llama-cpp, Alpaca, Saiga, Vicuna) - их можно запускать локально.