# Руководство по использованию Универсальный AI Помощник

## Быстрый старт

1. Введите ваш вопрос в верхнее текстовое поле.

2. (Необязательно) Загрузите файл документации (`.docx`, `.pdf`, `.txt`) или укажите URL документации.

3. (Необязательно) Для больших файлов включите "Расширенную обработку" и выберите размер фрагментов.

4. Нажмите "Спросить" или используйте Ctrl+Enter.

5. Ответ появится ниже, с форматированием Markdown и подсветкой кода.

---

## Поддерживаемые форматы файлов

- \*\*.docx\*\* — Microsoft Word

- \*\*.pdf\*\* — PDF-документы (только текстовые страницы)

- \*\*.txt\*\* — обычный текст

Максимальный размер файла: \*\*5 МБ\*\*

---

## Источники данных (приоритет)

1. \*\*Загруженный файл\*\* (docx/pdf/txt)

2. \*\*URL\*\* (будет скачан и проанализирован текст страницы, если файл не выбран)

3. \*\*data.docx\*\* (файл по умолчанию в корне проекта)

> \*\*Важно:\*\* Если указаны одновременно и файл, и URL — будет обработан только файл, а URL проигнорируется.

---

## Режимы обработки

### Стандартный режим

- Используется для небольших документов.

- Весь текст документа отправляется в модель целиком.

- Быстро, подходит для кратких инструкций, справок, небольших статей.

### Расширенная обработка

- Рекомендуется для больших документов (от 50 000 символов или >1 МБ).

- Документ разбивается на фрагменты (чанки) выбранного размера.

- Для поиска ответа используется семантический/векторный поиск:

- \*\*FAISS + Sentence Transformers\*\* — для больших документов

- \*\*TF-IDF\*\* — для небольших

- В ответе отображается, сколько фрагментов было использовано.

#### Как выбрать размер чанка?

- \*\*1000 символов\*\* — для очень структурированных или коротких абзацев

- \*\*2000 символов\*\* — универсальный вариант (по умолчанию)

- \*\*4000 символов\*\* — для длинных статей, справочников, технических мануалов

---

## Примеры вопросов

- Как создать образ Docker?

- Какой синтаксис у команды ... в моей документации?

- Как настроить сервис в предоставленном .docx?

- Как использовать функцию, описанную по ссылке?

- Какой параметр отвечает за ... в PDF?

---

## Рекомендации

- Для максимально точных ответов используйте расширенную обработку для больших файлов.

- Если документ содержит много таблиц/картинок (особенно PDF), результат может быть ограничен только текстовой частью.

- Для сложных технических вопросов лучше использовать документацию в виде .docx или .txt.

- Если ответ не найден — попробуйте переформулировать вопрос или загрузить более релевантный файл.

---

## Ограничения

- Только текстовые данные (PDF с картинками/сканами не поддерживаются)

- Максимальный размер файла — 5 МБ

- Ответ формируется на основе предоставленного контекста, не ищет в интернете

---

## Вопросы и поддержка

- Для обратной связи и багрепортов используйте GitHub Issues проекта.

Как работает помощник с эмбеддингами (простое объяснение)

1. Ввод вопроса

Пользователь вводит вопрос, например: "Как создать контейнер в Docker?"

2. Загрузка документа

Файл (.docx, .pdf, .txt) или URL загружается и разбивается на смысловые части — чанки.

3. Преобразование текста в эмбеддинги

Каждый фрагмент текста превращается в набор чисел — вектор, отражающий его смысл.

4. Преобразование вопроса в эмбеддинг

Вопрос тоже превращается в вектор с тем же алгоритмом.

5. Сравнение векторов

Система сравнивает вектор вопроса со всеми фрагментами документа и выбирает наиболее близкие по смыслу.

6. Формирование контекста для модели

Только эти релевантные фрагменты передаются в LLM (например, Gemini).

7. Ответ от модели

Модель возвращает ответ в формате Markdown с заголовками, кодом и списками.

Это работает как умный поиск по смыслу, а не просто по ключевым словам.

# Универсальный AI Помощник: Подробное руководство и описание

## Введение

Универсальный AI Помощник — это веб-приложение, разработанное на Flask и использующее мощь модели Gemini от Google. Его основная задача — предоставлять исчерпывающие ответы на ваши вопросы, основываясь на предоставленной документации. Вы можете загружать файлы в популярных форматах, указывать URL веб-страниц или использовать файл `data.docx`, который служит документацией по умолчанию.

## Ключевые возможности

- \*\*Многоформатная поддержка\*\*: Работает с файлами `.docx` (Microsoft Word), `.pdf` (только текстовые страницы) и `.txt` (обычный текст).

- \*\*Гибкие источники данных\*\*:

1. \*\*Загруженный файл\*\*: Наивысший приоритет. Если вы загружаете файл, он будет использован для ответов.

2. \*\*URL\*\*: Если файл не загружен, но указан URL, приложение скачает и проанализирует текстовое содержимое веб-страницы.

3. \*\*`data.docx`\*\*: Файл документации по умолчанию, расположенный в корне проекта. Используется, если не выбран ни файл, ни URL.

> \*\*Важно!\*\* При одновременном указании файла и URL, приоритет всегда отдается \*\*файлу\*\*. URL будет проигнорирован.

- \*\*Интеллектуальная обработка документов\*\*:

- \*\*Разбивка на чанки\*\*: Большие документы автоматически делятся на более мелкие, осмысленные фрагменты (чанки) для эффективной обработки.

- \*\*Семантический поиск\*\*: Приложение использует передовые методы семантического (смыслового) поиска для нахождения наиболее релевантных фрагментов документа, относящихся к вашему вопросу.

- \*\*Векторный поиск (FAISS)\*\*: Для очень больших документов применяется векторный поиск с использованием библиотеки FAISS и моделей `Sentence Transformers` (например, `all-MiniLM-L6-v2`), что позволяет быстро находить смысловую близость между вопросом и частями текста.

- \*\*TF-IDF\*\*: Для документов меньшего размера используется классический подход TF-IDF для определения релевантности.

- \*\*Интерактивный интерфейс\*\*:

- \*\*Markdown и подсветка кода\*\*: Ответы форматируются с использованием Markdown (заголовки, списки, выделение) и автоматически подсвечивают синтаксис для различных языков программирования в блоках кода.

- \*\*Темы оформления\*\*: Поддерживается светлая и тёмная темы с возможностью автоматического определения системных настроек и ручного переключения.

- \*\*Адаптивный дизайн\*\*: Интерфейс корректно отображается и удобен для использования на мобильных устройствах и планшетах.

- \*\*Настраиваемость\*\*:

- \*\*Выбор режима обработки\*\*:

- \*\*Стандартный режим\*\*: Идеален для небольших документов. Весь текст документа передается модели целиком. Обеспечивает быстрые ответы.

- \*\*Расширенная обработка\*\*: Рекомендуется для объемных документов (например, более 50 000 символов или >1 МБ). Активирует разбивку на чанки и семантический/векторный поиск.

- \*\*Настройка размера чанков\*\*: В режиме расширенной обработки можно выбрать оптимальный размер фрагментов (1000, 2000 или 4000 символов) в зависимости от структуры и типа документа.

- \*\*Модель Gemini\*\*: Используемую модель Gemini можно настроить через переменные окружения (`GEMINI\_MODEL`). По умолчанию используется `models/gemini-2.5-flash-preview-04-17`.

- \*\*Обработка файлов и URL\*\*:

- \*\*Ограничение размера файла\*\*: Максимальный размер загружаемого файла — 5 МБ.

- \*\*Проверка URL\*\*: Перед обработкой URL проверяется на корректность.

- \*\*Улучшенная обработка ошибок\*\*:

- Приложение предоставляет понятные сообщения об ошибках, если возникают проблемы с загрузкой или обработкой файлов/URL (например, некорректный URL, недоступность ресурса, неподдерживаемый формат файла, превышение размера файла).

- В случае ошибки с указанным источником, приложение автоматически переключится на использование документации по умолчанию (`data.docx`) и уведомит об этом пользователя.

## Как это работает (принцип расширенной обработки)

1. \*\*Ввод вопроса\*\*: Пользователь задает вопрос (например, "Как настроить аутентификацию в системе?").

2. \*\*Загрузка и подготовка документа\*\*:

- Выбирается источник данных (файл, URL или `data.docx`).

- Текст документа извлекается и очищается.

- Если включена расширенная обработка, текст разбивается на чанки (смысловые фрагменты).

3. \*\*Создание векторных представлений (эмбеддингов)\*\*:

- Каждый чанк текста преобразуется в числовой вектор (эмбеддинг) с помощью модели `Sentence Transformer`. Этот вектор отражает семантическое значение чанка.

- Вопрос пользователя также преобразуется в аналогичный вектор.

4. \*\*Поиск релевантных чанков\*\*:

- Система сравнивает вектор вопроса с векторами всех чанков документа, используя FAISS (для больших документов) или TF-IDF (для небольших).

- Выбираются наиболее близкие по смыслу чанки (`top\_k`).

5. \*\*Формирование контекста для LLM\*\*:

- Отобранные релевантные чанки объединяются и формируют контекст, который будет передан большой языковой модели (LLM), такой как Gemini.

6. \*\*Генерация ответа\*\*:

- LLM анализирует вопрос пользователя и предоставленный контекст из документа.

- Модель генерирует развернутый ответ, стараясь использовать только информацию из документа.

7. \*\*Отображение ответа\*\*:

- Ответ возвращается пользователю в отформатированном виде (Markdown, подсветка кода).

- Может отображаться информация о количестве использованных и общем числе фрагментов.

Этот подход позволяет модели сосредоточиться на наиболее релевантных частях документа, что повышает точность ответов и эффективность обработки больших объемов информации.

## Быстрый старт использования

1. \*\*Введите ваш вопрос\*\*: Напишите свой вопрос в текстовое поле в верхней части страницы.

2. \*\*Укажите источник данных (необязательно)\*\*:

- \*\*Загрузите файл\*\*: Нажмите на кнопку "Файл" и выберите документ (`.docx`, `.pdf`, `.txt`) со своего компьютера.

- \*\*Укажите URL\*\*: Введите URL веб-страницы с документацией в соответствующее поле.

3. \*\*Выберите режим обработки (необязательно)\*\*:

- Для небольших документов стандартный режим обычно достаточен.

- Для больших документов рекомендуется включить "Расширенную обработку" и, при необходимости, выбрать размер чанков (1000, 2000 или 4000 символов). Опция "Средние части (2000 символов)" является универсальным вариантом по умолчанию.

4. \*\*Отправьте запрос\*\*: Нажмите кнопку "Спросить" или используйте комбинацию клавиш `Ctrl+Enter`.

5. \*\*Получите ответ\*\*: Ответ от AI Помощника появится в блоке ниже, отформатированный для удобного чтения.

## Рекомендации по использованию

- \*\*Для больших документов\*\*: Всегда используйте "Расширенную обработку" для получения наиболее точных и релеванттных ответов. Это особенно важно для технических мануалов, спецификаций или объемных статей.

- \*\*Структура документа\*\*: Четко структурированные документы с логичными абзацами и заголовками обычно обрабатываются лучше.

- \*\*Формат файла\*\*: Для сложной технической документации или документов с таблицами и форматированием, `.docx` или `.txt` могут дать лучший результат при извлечении текста, чем `.pdf`, особенно если PDF содержит много изображений или сложную верстку.

- \*\*Переформулировка вопроса\*\*: Если вы не получили удовлетворительный ответ, попробуйте переформулировать свой вопрос или разбить сложный вопрос на несколько более простых.

- \*\*Выбор размера чанка\*\*:

- \*\*1000 символов\*\*: Подходит для документов с очень короткими, четко разделенными информационными блоками или когда требуется высокая детализация ответа из небольших фрагментов.

- \*\*2000 символов\*\*: Универсальный вариант, хорошо работает для большинства документов.

- \*\*4000 символов\*\*: Используйте для очень длинных статей, справочников, где информация может быть распределена по большим разделам текста.

## Ограничения

- \*\*Только текстовые данные\*\*: Приложение обрабатывает только текстовую информацию. Изображения, сканы в PDF или встроенные объекты не анализируются.

- \*\*Качество извлечения текста из PDF\*\*: Качество текста, извлеченного из PDF, может варьироваться в зависимости от структуры PDF-файла. PDF, созданные из текстовых редакторов, обрабатываются лучше, чем сканированные PDF.

- \*\*Максимальный размер файла\*\*: Ограничение на размер загружаемого файла — 5 МБ.

- \*\*Контекст ответа\*\*: Ответы генерируются исключительно на основе предоставленной документации. Приложение не ищет информацию в интернете и не использует свои "общие знания" вне контекста документа.

- \*\*Производительность\*\*: Обработка очень больших файлов в режиме "Расширенная обработка" может занять некоторое время, особенно при первой загрузке модели эмбеддингов.

## Установка и запуск (для разработчиков)

1. \*\*Клонируйте репозиторий\*\*:

```bash

git clone <URL-репозитория>

cd <название-папки-проекта>

```

2. \*\*Установите зависимости\*\*: Убедитесь, что у вас установлен Python 3.8+ и pip.

```bash

pip install -r requirements.txt

```

3. \*\*Настройте API ключ Gemini\*\*:

- Скопируйте файл `.env.example` в `.env`: `cp .env.example .env`

- Откройте файл `.env` и вставьте свой `GEMINI\_API\_KEY`.

- (Опционально) Вы можете указать другую модель Gemini в переменной `GEMINI\_MODEL`.

4. \*\*Документация по умолчанию (необязательно)\*\*: Поместите файл `data.docx` с вашей основной документацией в корень проекта.

5. \*\*Запустите приложение\*\*:

```bash

python app.py

```

Приложение будет доступно по адресу `http://127.0.0.1:5000/`.

## Зависимости проекта

- Flask: микрофреймворк для веб-приложения.

- python-docx: для чтения `.docx` файлов.

- google-generativeai: официальная библиотека Google для работы с API Gemini.

- gunicorn: WSGI HTTP-сервер для запуска Flask-приложений в production.

- python-dotenv: для управления переменными окружения.

- requests: для выполнения HTTP-запросов (загрузка URL).

- PyPDF2: для извлечения текста из `.pdf` файлов.

- langchain, langchain-text-splitters: для разбивки текста на чанки.

- faiss-cpu: для эффективного векторного поиска.

- sentence-transformers: для создания текстовых эмбеддингов.

- scikit-learn: для TF-IDF и вычисления косинусного сходства.

- numpy: для работы с числовыми массивами (эмбеддингами).

- nltk, tiktoken: вспомогательные библиотеки для обработки текста (могут требоваться для некоторых моделей `SentenceTransformer` или `langchain`).

- chromadb: (Хотя указан в README, текущая реализация в `app.py` использует FAISS и TF-IDF напрямую, а не ChromaDB для векторного хранения).

## Вопросы и поддержка

Для предложений, сообщений об ошибках или вопросов по использованию, пожалуйста, создавайте "Issue" в GitHub-репозитории проекта.