Модуль речевого ввода данных с LLM-корректором

Информация, необходимая для установки программного обеспечения

Работа на здоровье

INTERIN
технологии

2026 г.

Модуль речевого ввода данных с LLM-корректором

Информация, необходимая для установки программного обеспечения

Контактная информация

Группа компаний «Интерин»
Web: www.interin.ru
E-mail: info@interin.ru
Тел: +7 (495) 2208235

ОБЩАЯ ИНФОРМАЦИЯ

При штатной работе в составе МИС программа устанавливается на сервер медицинской организации, работает в изолированном контуре, не требует постоянного доступа к инфраструктуре разработчика.

СОСТАВ ДИСТРИБУТИВА

1. Дистрибутив программного обеспечения (далее – Модуль) поставляется в виде архива (например, .tar.gz или .zip), содержащего следующие компоненты:

Компонент	Описание
server/	Серверная часть Модуля (Python-приложение)
client/	Клиентская часть (веб-интерфейс)
models/	Предустановленные модели (Whisper, Vosk, Qwen) – опционально
scripts/	Скрипты для установки и настройки
config/	Шаблоны конфигурационных файлов
requirements.txt	Зависимости Python
docker-compose.yml	Файл для развертывания через Docker
INSTALL.sh	Установочный скрипт для Debian/Ubuntu

АППАРАТНЫЕ И ПРОГРАММНЫЕ ТРЕБОВАНИЯ

2. Аппаратные требования

2.1 Сервер

Компонент	Минимальные требования	Рекомендуемые требования
CPU	Intel Xeon 2.0 ГГц, 8 ядер	Intel Xeon 2.5 ГГц, 16+ ядер
GPU	RTX 4090 24 ГБ	RTX 4090 / A100 24–48 ГБ
RAM	32 ГБ	64 ГБ (для комфортной работы с Qwen2.5-32B)
Дисковое пространство	500 ГБ (RAID 10)	1 ТБ (SSD, RAID 10)
Сеть	100 Мбит/сек	1 Гбит/сек

2.2 Рабочая станция пользователя

Компонент	Требование
CPU	от 2 ГГц
RAM	от 4 ГБ
Дисковое пространство	от 80 ГБ
Видеокарта	поддержка разрешения 1280x1024 и выше
Монитор	от 19.0" LCD
Сеть	100 Мбит/сек

3. Программные требования

3.1 Операционная система (сервер)

- Debian 12 и выше
- Ubuntu 24.04 и выше

3.2 Клиентская часть (рабочая станция)

Веб-браузер актуальной версии (один из списка):

- Яндекс.Браузер
- Спутник
- Microsoft Edge
- Google Chrome
- Mozilla Firefox
- Opera

3.3 Системные зависимости (сервер)

При установке автоматически проверяются и устанавливаются:

- Python 3.10+
- Docker (при использовании контейнеризации)
- Docker Compose
- ffmpeg
- CUDA Toolkit (для GPU)
- NVIDIA Container Toolkit (для Docker с GPU)

4. Сторонние компоненты (библиотеки)

Модуль использует следующие библиотеки с разрешительными лицензиями (MIT, Apache 2.0, BSD):

Библиотека	Версия	Лицензия
aiofiles	25.1.0	MIT
aiohttp	3.13.3	Apache 2.0
faster-whisper	1.2.1	MIT
langchain-core	1.2.16	MIT
onnxruntime	1.24.2	MIT
pydantic	2.12.5	MIT
torch	2.10.0	BSD 3-Clause
vosk	0.3.45	Apache 2.0
websockets	16.0	BSD 3-Clause

Полный список приведён в документации (Модуль речевого ввода данных с LLM-коррек-тором. Руководство пользователя).

5. Поддерживаемые модели ИИ

Модель	Тип	Лицензия
Qwen2.5-32B-Instruct (4–8 бит)	LLM	Apache 2.0
Vosk (ru-0.22 и др.)	ASR	Apache 2.0
Whisper (small, large-v3 и др.)	ASR	MIT

ПОРЯДОК УСТАНОВКИ

Порядок установки

1) Установка на сервер (Debian/Ubuntu)

Способ 1: Автоматическая установка

bash

```
tar -xzf interin-speech-module.tar.gz
cd interin-speech-module
sudo ./INSTALL.sh
```

Скрипт выполнит:

- проверку системных требований
- установку зависимостей (apt, pip)
- настройку GPU (CUDA)
- развертывание сервисов через systemd или Docker

Способ 2: Установка через Docker

bash

```
docker-compose up -d
```

2) Настройка конфигурации

После установки необходимо отредактировать файл .env или config.yaml:

yaml

```
asr_service: "whisper"   # whisper / vosk
llm_model: "qwen"        # qwen / другие
llm_endpoint: "http://localhost:8000"
gpu_enabled: true
template_path: "./templates"
```

3) Установка клиентской части

Клиентская часть не требует установки – доступна через веб-браузер по адресу:

text

```
http://<сервер_IP>:8080
```

ИНТЕГРАЦИЯ С МИС

Для интеграции с медицинской информационной системой (МИС) необходимо:

1) Настроить API-взаимодействие (REST / WebSocket) между Модулем и МИС.
2) Передать шаблоны медицинских документов в формате JSON/XML.
3) Настроить единую систему аутентификации (при необходимости).
4) В МИС добавить кнопку вызова речевого ввода для соответствующих форм.

ПРОВЕРКА РАБОТОСПОСОБНОСТИ

После установки рекомендуется выполнить тестовый сценарий:

1) Открыть веб-интерфейс.
2) Включить микрофон.
3) Произнести тестовую фразу (например: «Правая почка 10x5 см, паренхима не изменена»).
4) Выключить микрофон.
5) Нажать «Отправить в LLM на коррекцию».
6) Убедиться в получении структурированного документа.

УСТРАНЕНИЕ НЕПОЛАДОК ПРИ УСТАНОВКЕ

Проблема	Возможное решение
Не хватает GPU памяти	Уменьшить размер пакета аудио или использовать меньшую модель
Ошибка CUDA	Установить CUDA Toolkit и NVIDIA драйвер
Не запускается веб-интерфейс	Проверить порт 8080: `netstat -tulpn | grep 8080`
Модель не загружается	Проверить доступ к интернету (для скачивания) или смонтировать локальный кэш моделей