



Разработка ИИ-агента с возможностью генерации речи

Сальников С. С. Ячmeneва Н. Н.
16-18 апреля 2026



IBM. | **Deeprgram**

cabolo®

WHISPP 

Предметная область

- Задача: разработка системы для реализации взаимодействия с большими языковыми моделями с возможностью копирования голоса

Приоритеты



Узкая специализация

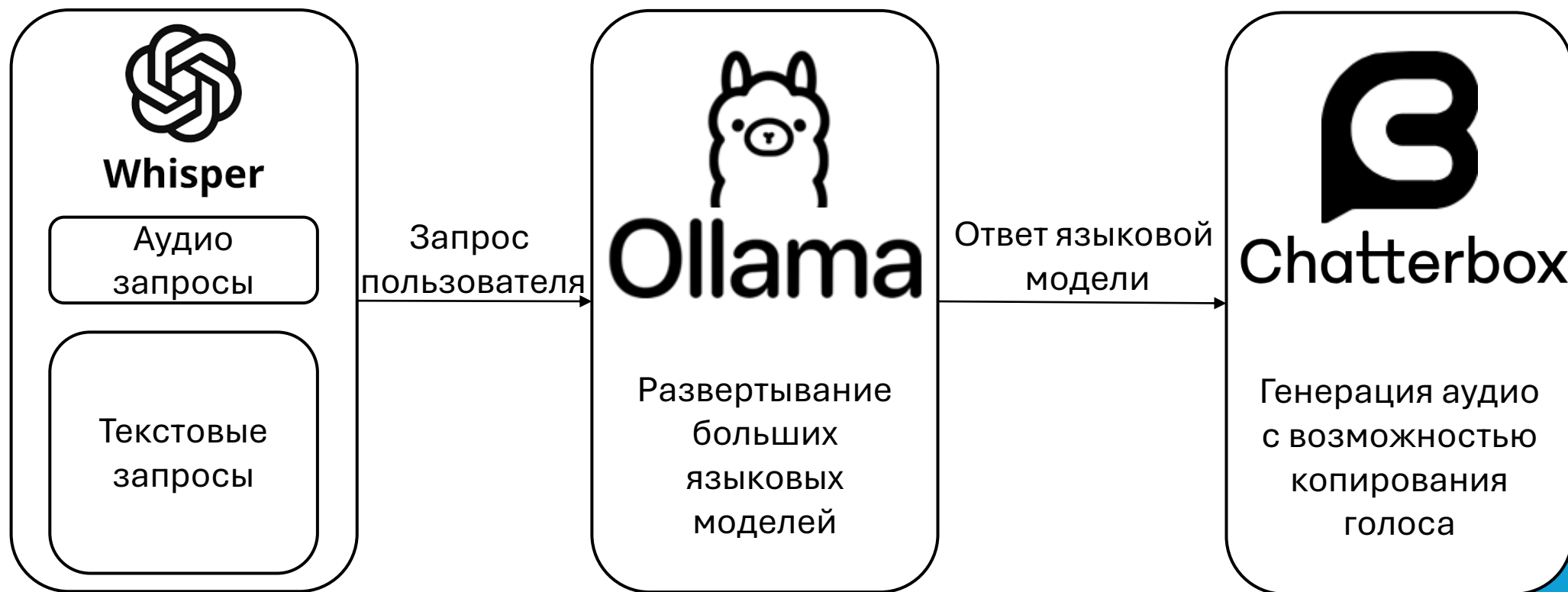


Работа в реальном
времени

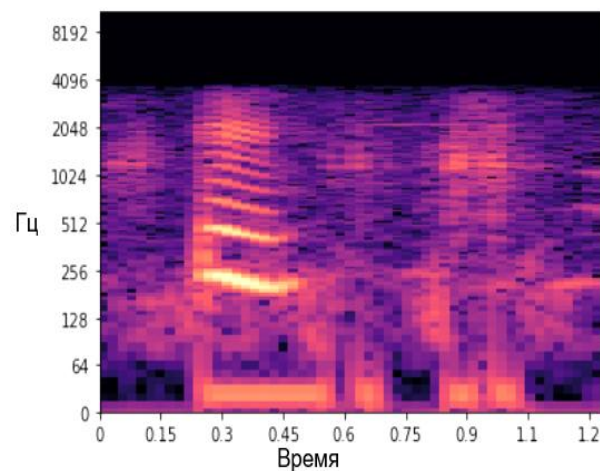


Приватность

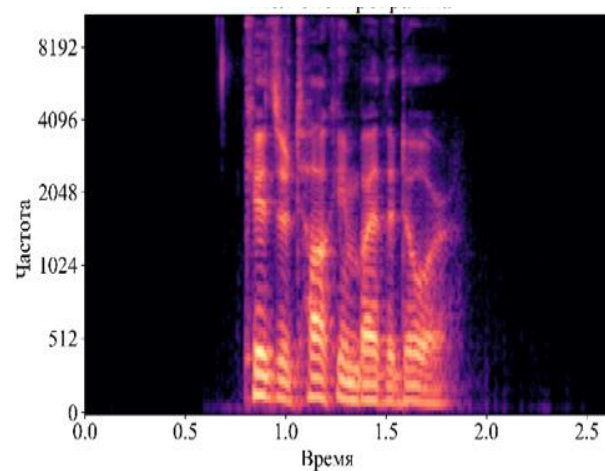
Структура



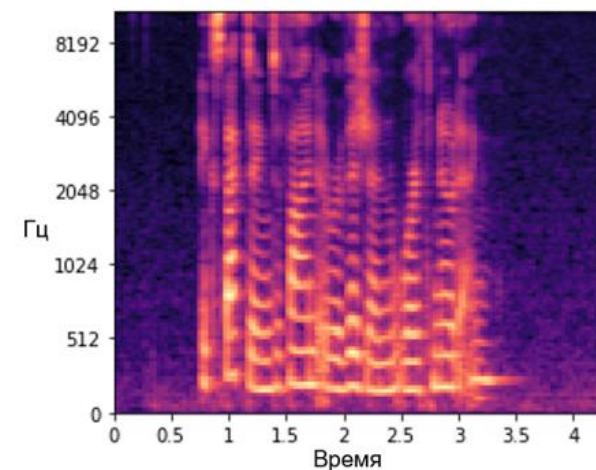
Спектрограммы



Стандартная спектрограмма



Mel-спектрограмма



Log-Mel спектрограмма

Whisper AI

Сегментирование аудио:

разбиение на сегменты по 30 с. в 16 кГц

Кодировка:

преобразование спектрограмм в векторы с информацией об аудио

Декодировка:

авторегрессия текста по токенам контекста

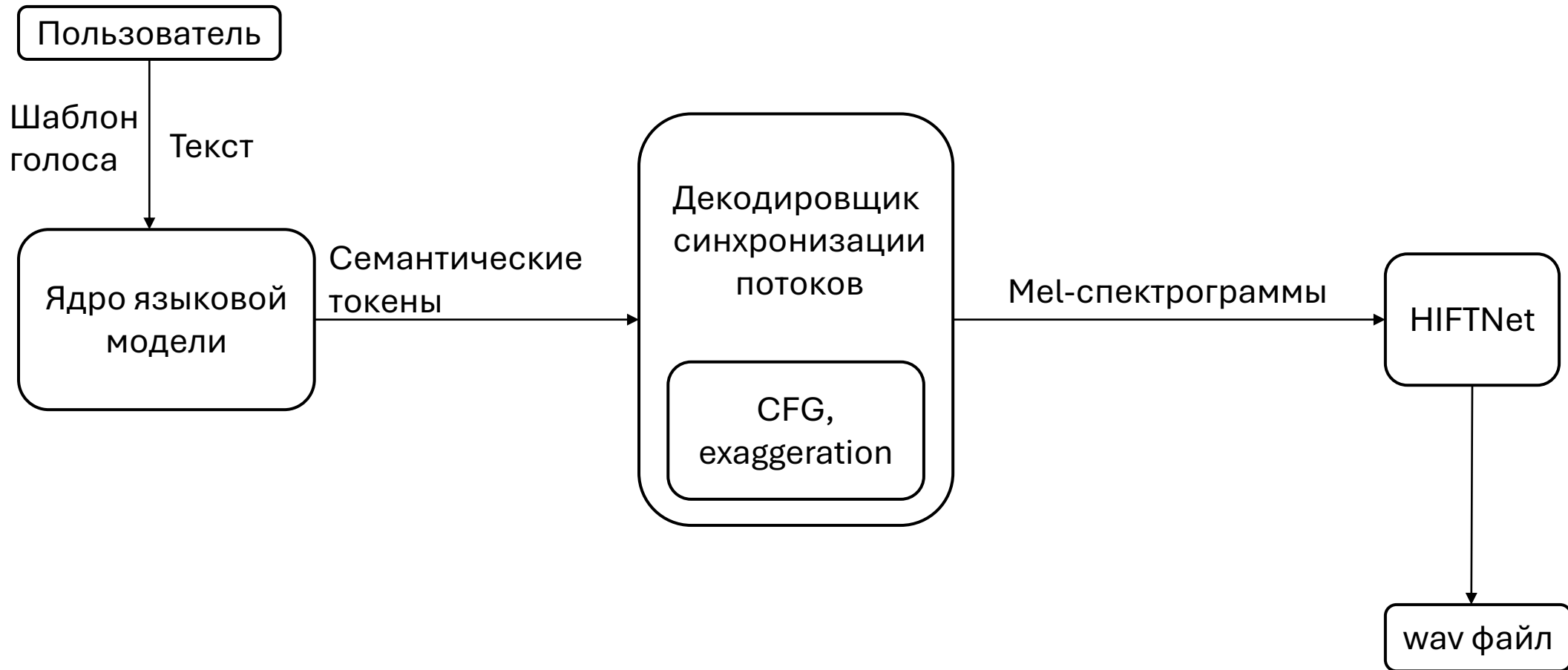
Ollama

- Инструмент для развертывания языковых моделей
Позволяет запускать ядра языковых моделей, как локально, так и из облака сервиса

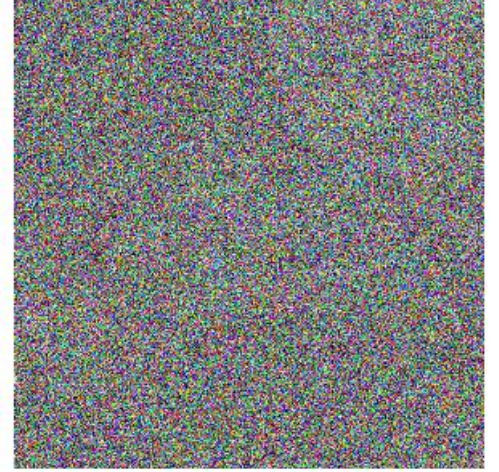


Ollama

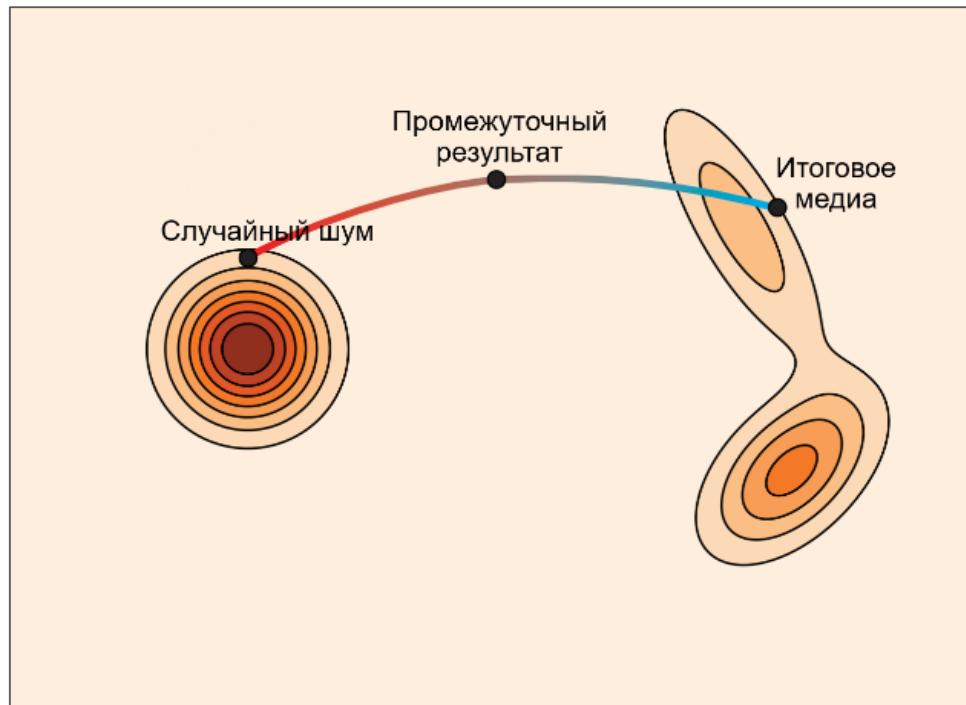
Chatterbox TTS



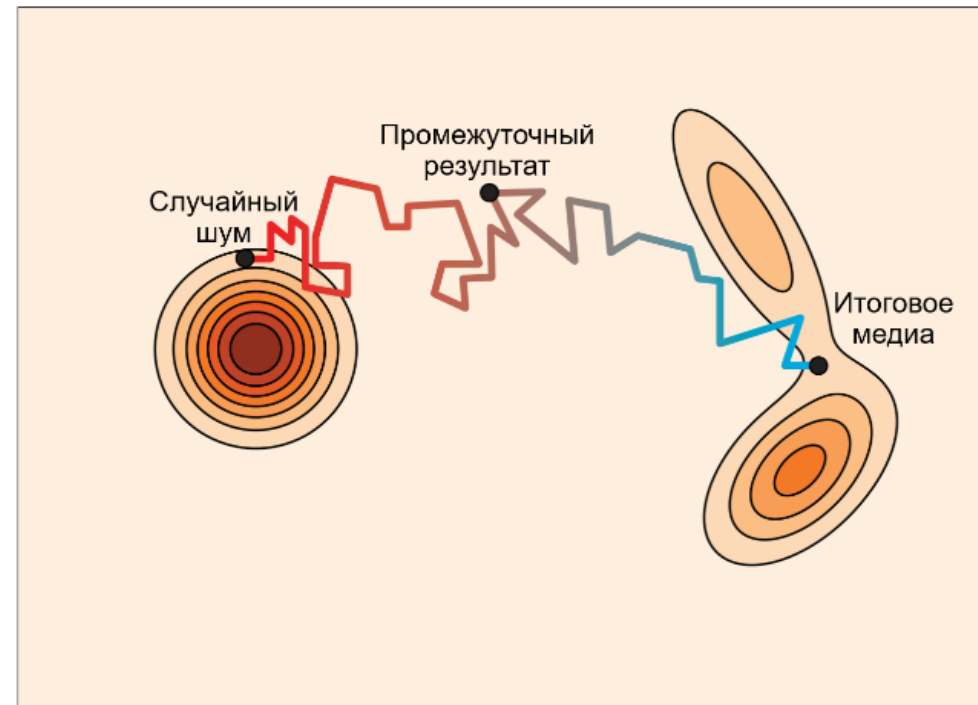
Синхронизация потоков (flow matching)



Синхронизация потоков (flow matching)



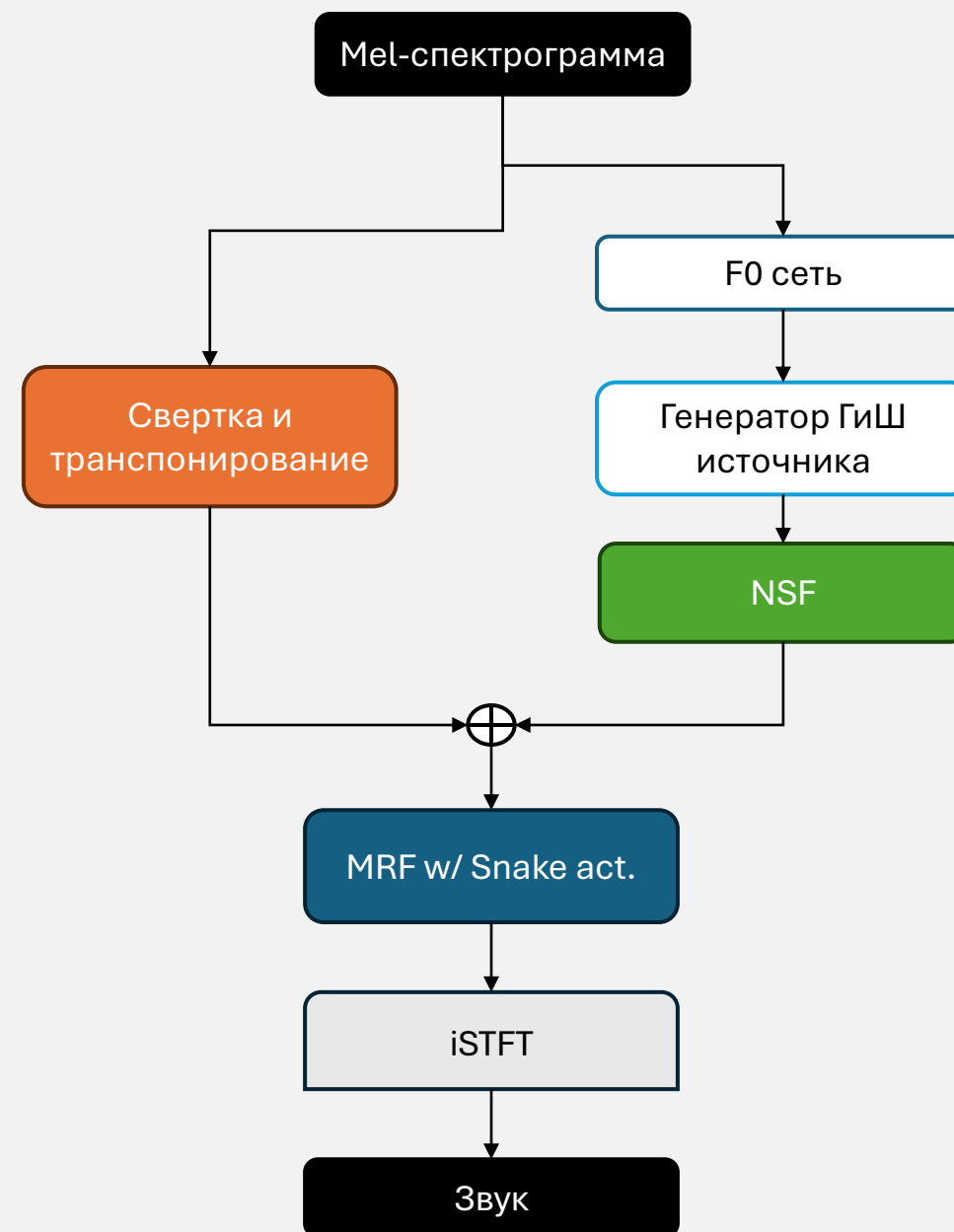
Поток



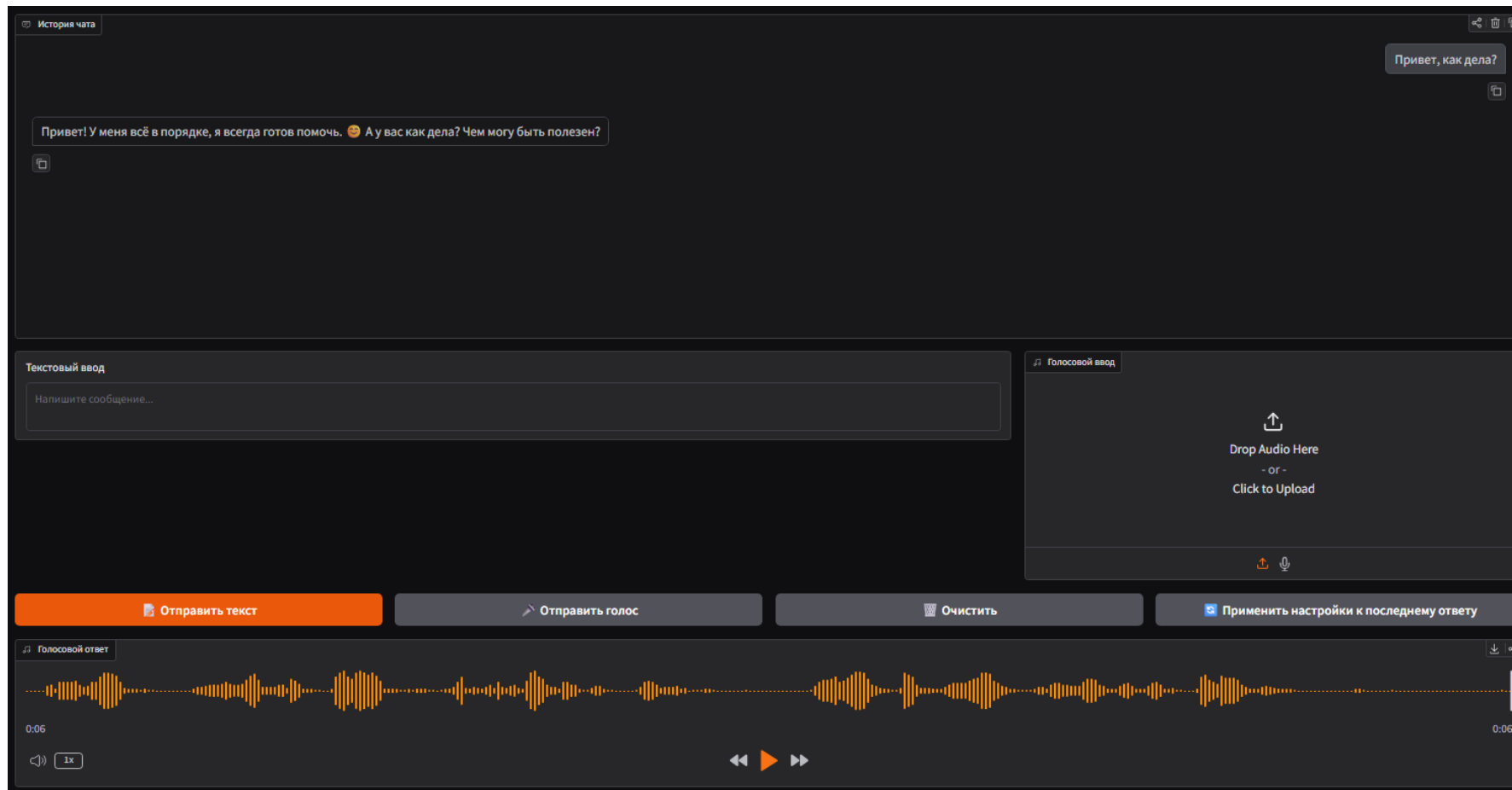
Диффузия

HiFTNet

- F0 сеть обучена генерации фундаментальных частот
- Фундаментальная частота преобразуется в гармоническо-шумовой источник
- Нейронный фильтр источника генерирует звук из источника
- Мульти-рецепционное поле добавляет детали ($snake(x) = x + \frac{1}{\alpha} \sin^2(\alpha x)$)
- Получаем чистый звук с помощью ОКвПФ



Диалоговое окно приложения



Настройка параметров генерации

Настройки голоса

☐

Использовать мой голос как референс

Если включено и отправлен голос, TTS скопирует ваш голос (приоритет выше остальных)

☐

Использовать референс голоса

Если включено, TTS будет имитировать голос из файла ниже

🔊 Файл референса (для TTS)

📶

Drop Audio Here

- or -

Click to Upload

CFG Scale

Влияние классификатора (1.0 = нейтрально, выше = выразительнее)

0.5

3

Exaggeration

Усиление эмоций/артикуляции (1.0 = нормально)

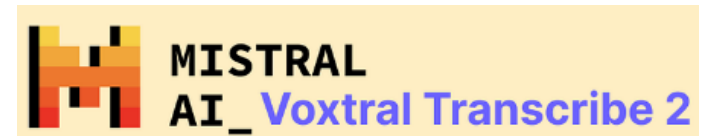
0.5

2.5

More options

Итоги и планы

- Реализовано приложение для голосового общения с большими языковыми моделями со стилизацией ответов.
- Переход на сервисы по типу Open WebUI для создания полноценного ИИ-агента, вместо приложения, утилизирующего ИИ API.
- Смена архитектуры транскрибирующей модели на оптимизированные под работу в реальном времени, по типу Voxtral Transcribe 2 от Mistral.





Спасибо за внимание