Повышение точности субтитров в художественных фильмах: технологии и опыт

Шатнёв Дмитрий, 2024

О себе

- Инженер машинного обучения в Inventos
- Студент ОГУ им. И.С. Тургенева
- Сферы научных интересов:
  - Обработка естественного языка
  - Речевые технологии

План доклада

Webcaster - отправная точка

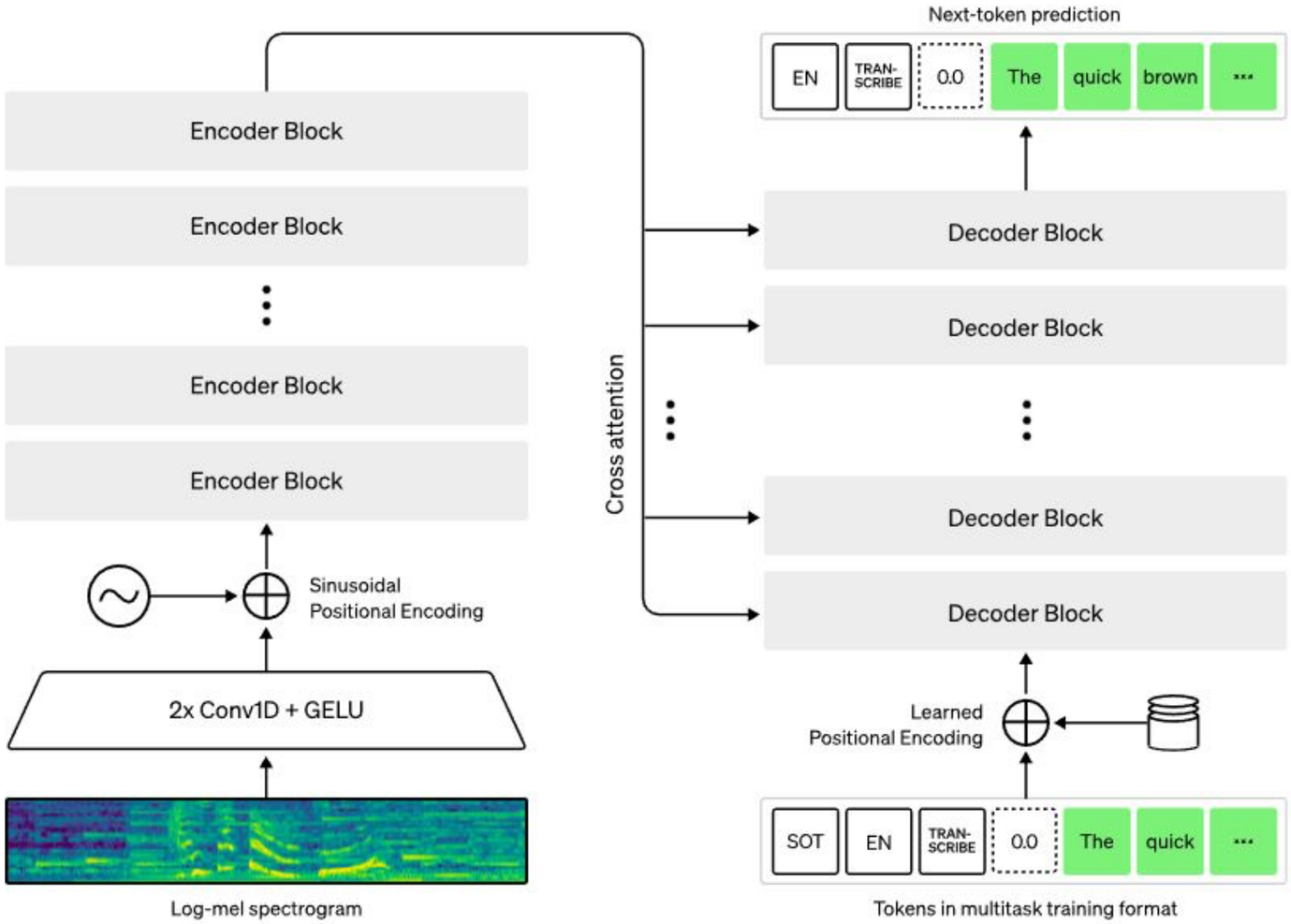
https://webcaster.pro/

Особенности контента

- Существенная длительность
- Сложная звуковая картина
- Мультиязычность
- Аудиовизуальная корреляция

Action Vectors by Vecteezy

Модель Whisper



https://github.com/openai/whisper

Первая итерация

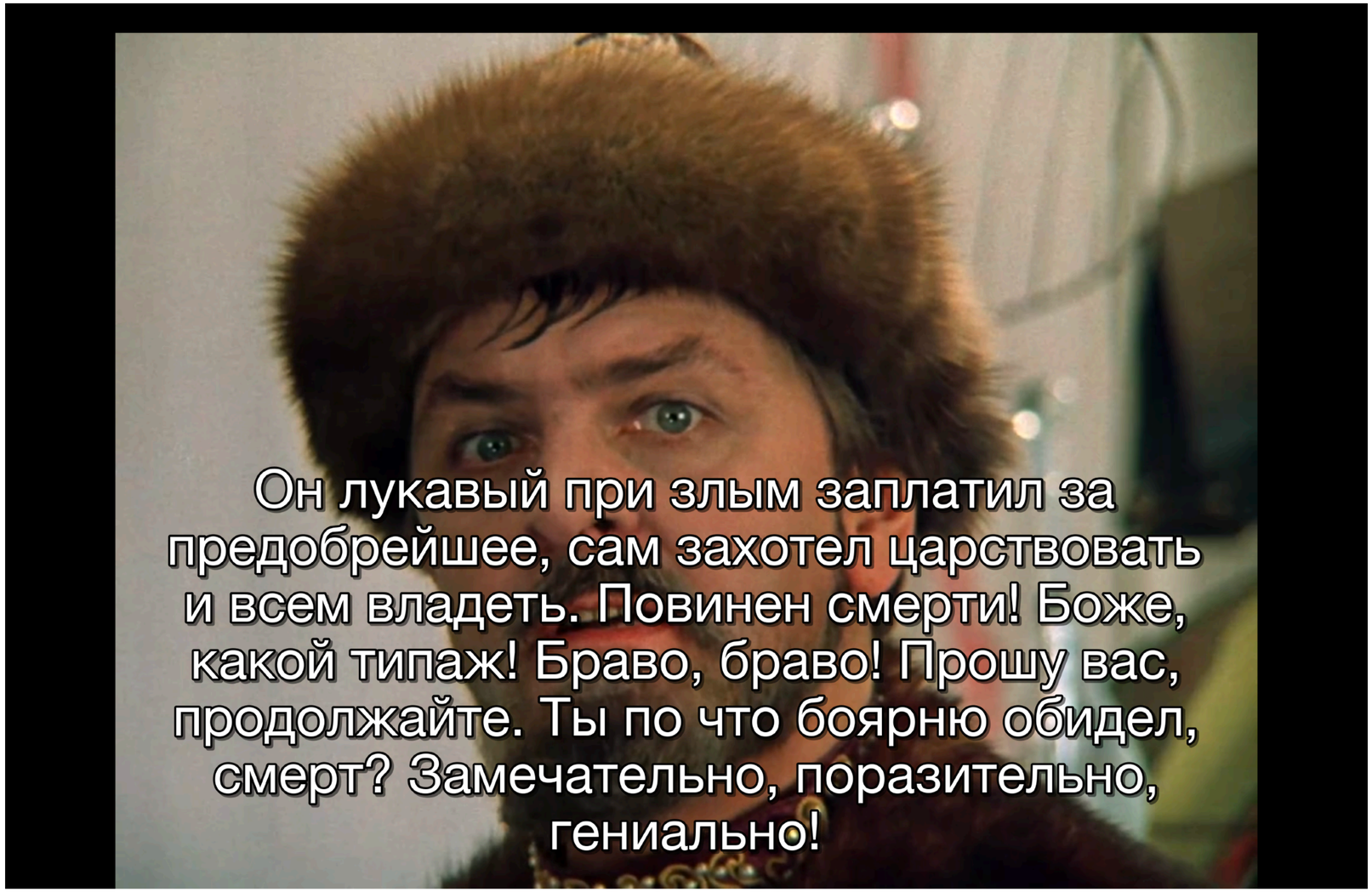
Использование Whisper из коробки:

- Максимально простой пайплайн
- Минимум усилий

Bad case #1

https://www.mosfilm.ru/

Bad case #2



https://www.mosfilm.ru/

Первые проблемы

- Аномалии распознавания
- Рассинхронизация по временным меткам
- Большой объем текста в блоке субтитров
- Сложность объективной оценки

Основная метрика качества

$$\text{Word Error Rate (WER)} = \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions}}{\text{Number of words in reference}}$$

Accuracy = 100% - WER

Эталон: Жизнь как коробка шоколадных конфет.

Распознано: Жизнь коровка шоколадных конфет. Пока!

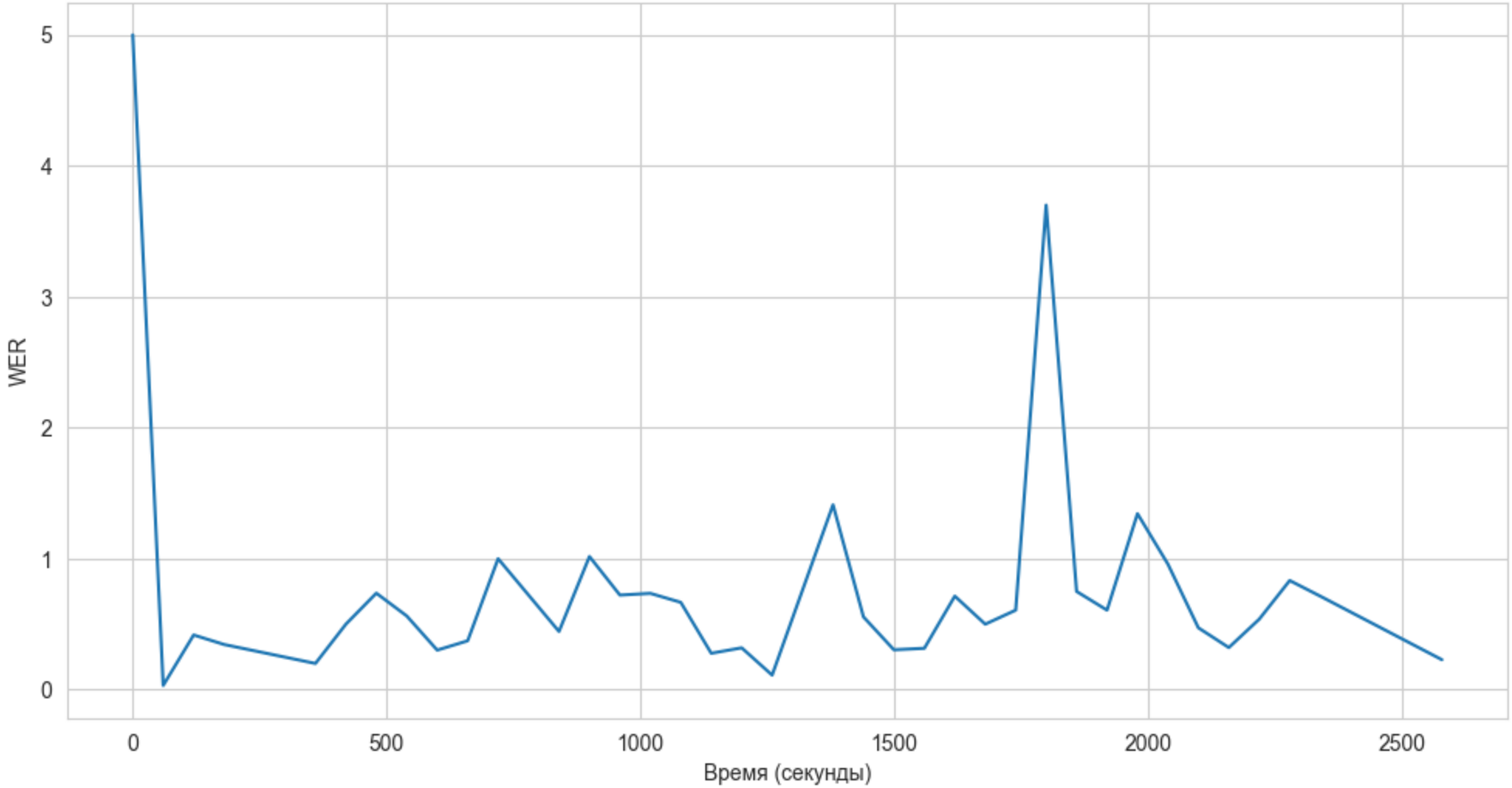
Дополнительные метрики качества

- CER (Character Error Rate)
- WIP (Word Information Preserved)
- WIL (Word Information Lost)

Metrics Vectors by Vecteezy

Локализация падения метрик

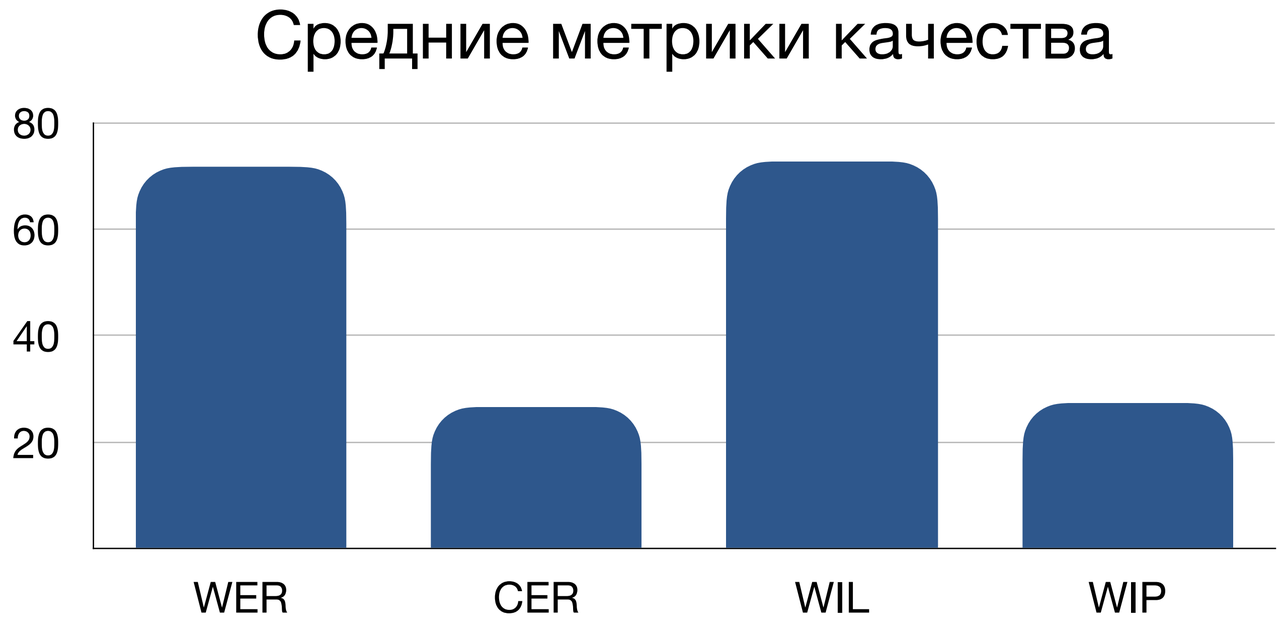
Window: 1 min.
Step: 1 min.
Audio duration: 45 min.



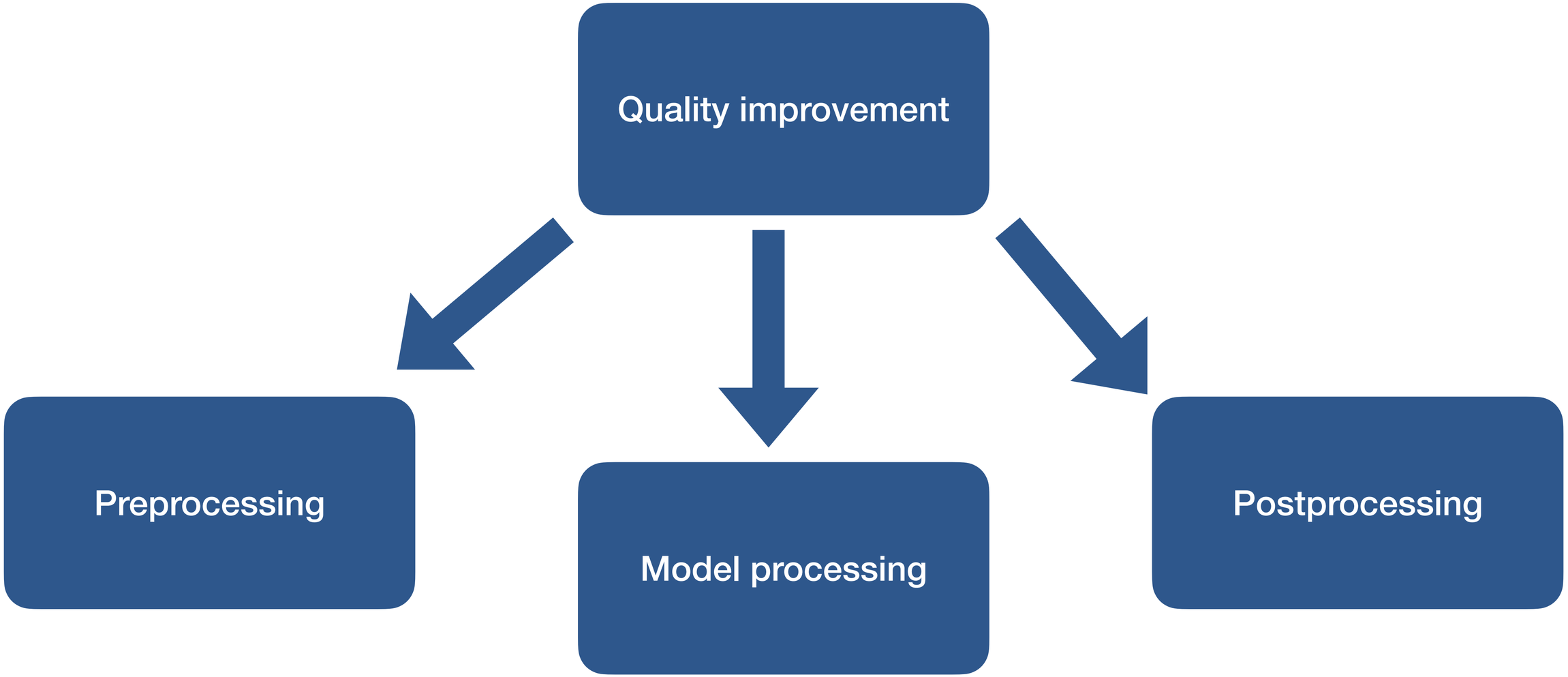
Промежуточные выводы

Текущие показатели свидетельствуют о **низкой** точности распознавания.

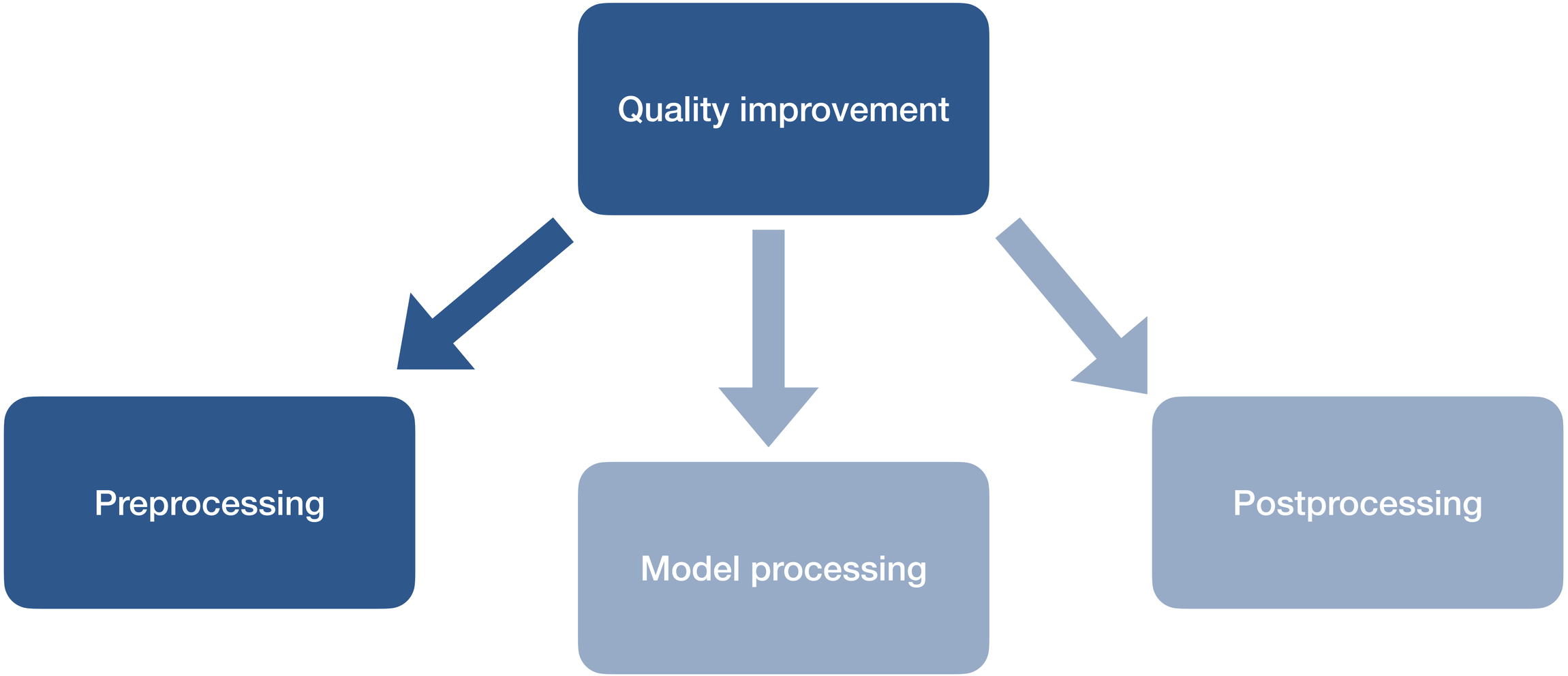
Необходимо комплексно оптимизировать систему, для достижения приемлемого качества субтитров.



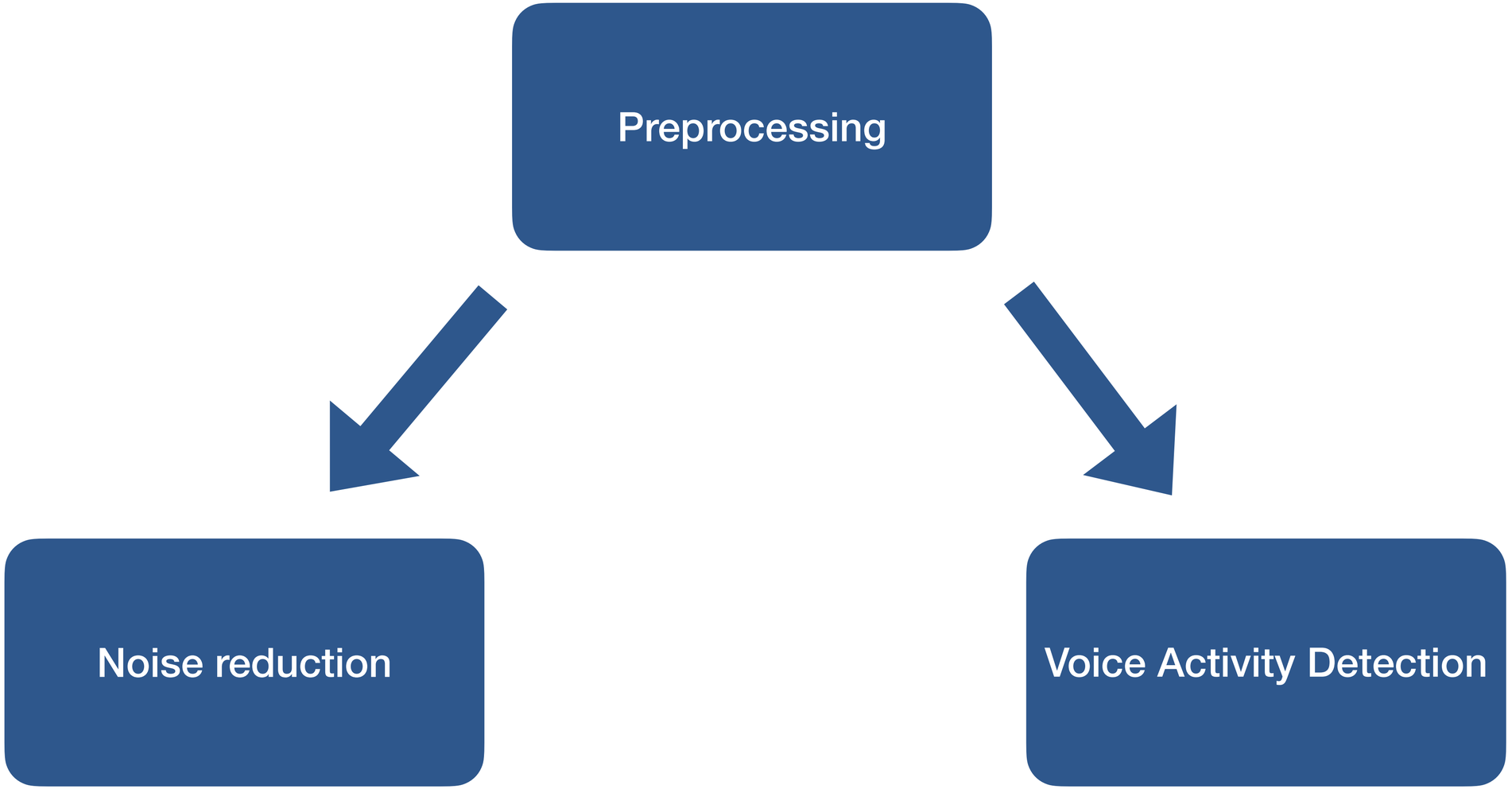
Способы улучшения качества



Препроцессинг



Вариации препроцессинга



Шумоподавление

В теории:

Улучшает качество входного аудио

Облегчает выделение речевого сигнала

Уменьшает количество ошибок распознавания

На практике:

Удаляет важные звуковые детали

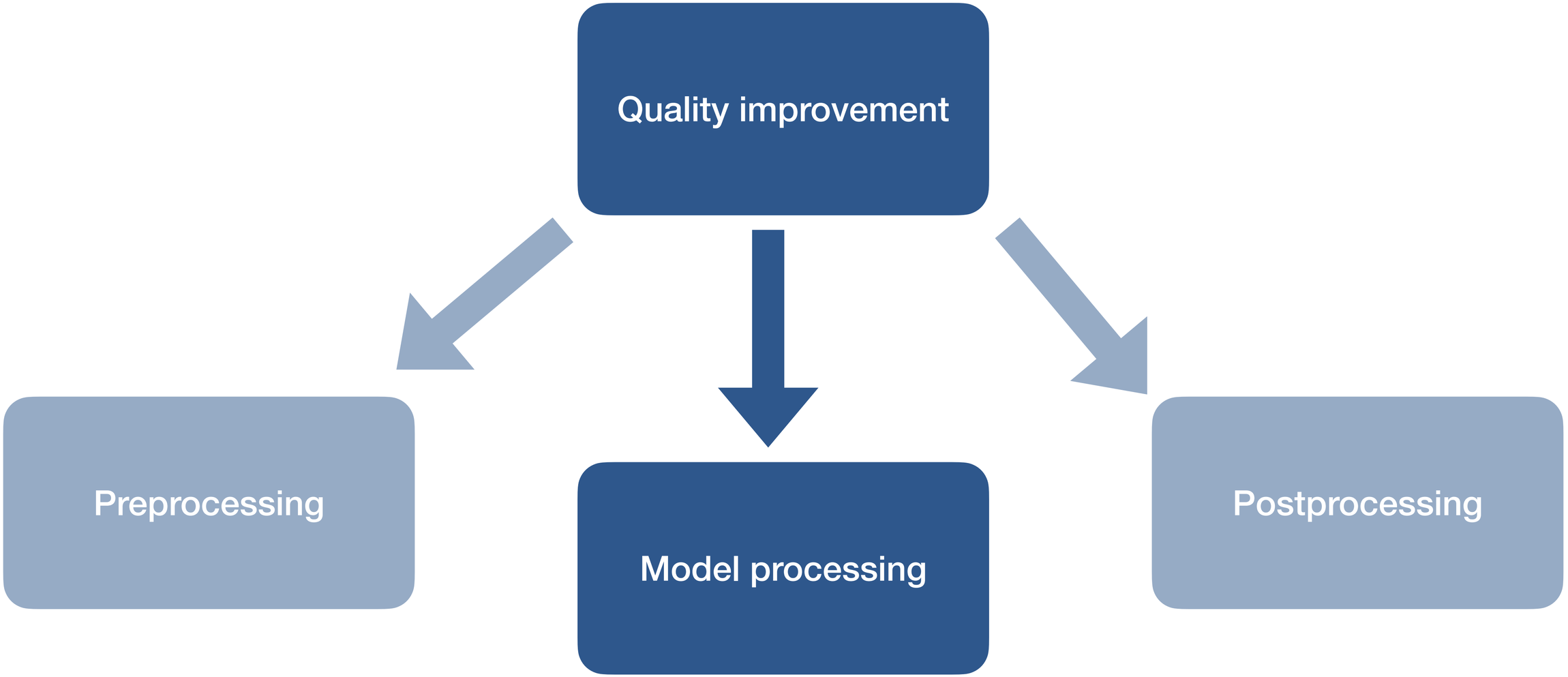
Ухудшает распознавание в сценах с наложением голосов

Требует тонкой настройки для каждого фильма

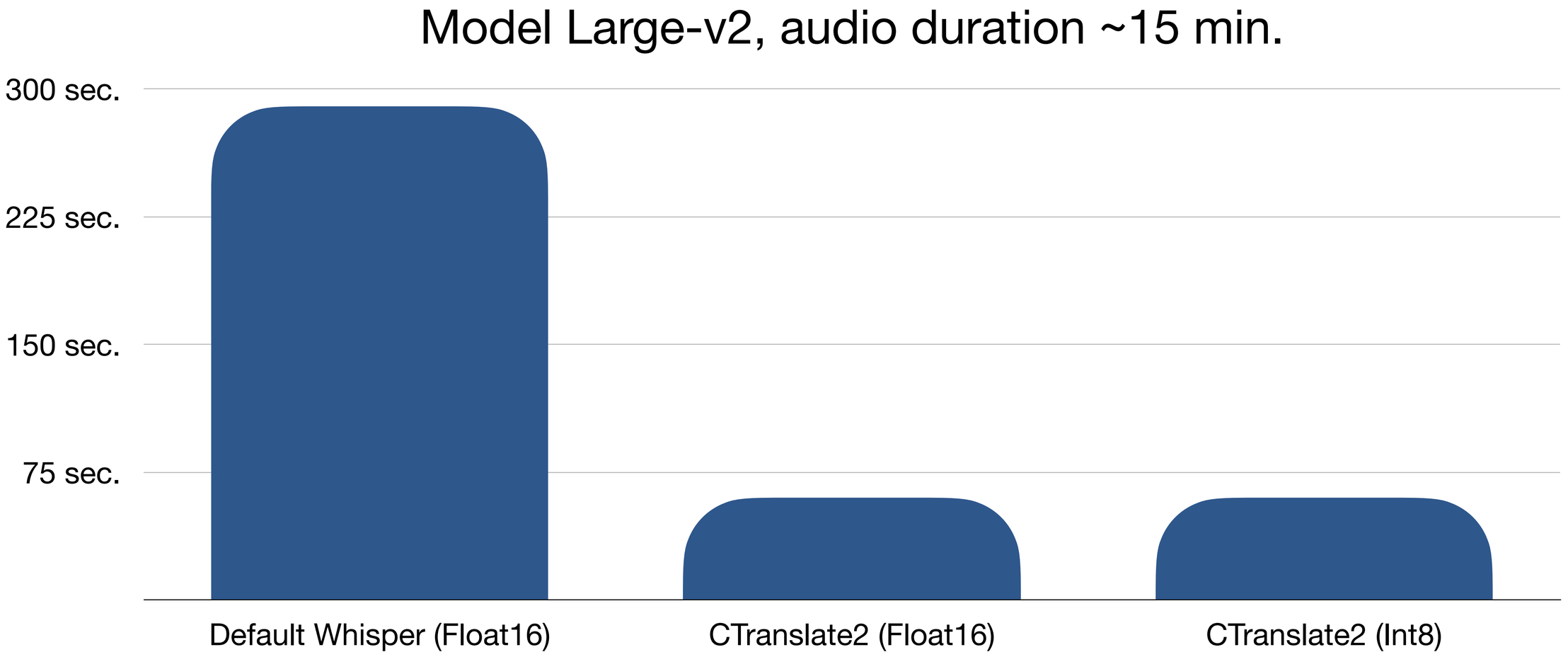
Voice Activity Detection

- Сосредоточивает вычислительные ресурсы только на значимых сегментах аудио
- Минимизирует аномалии на фоновых шумах и музыке
- Повышает робастность системы к сложным акустическим условиям
- Упрощает интеграцию с системами анализа и идентификации говорящего

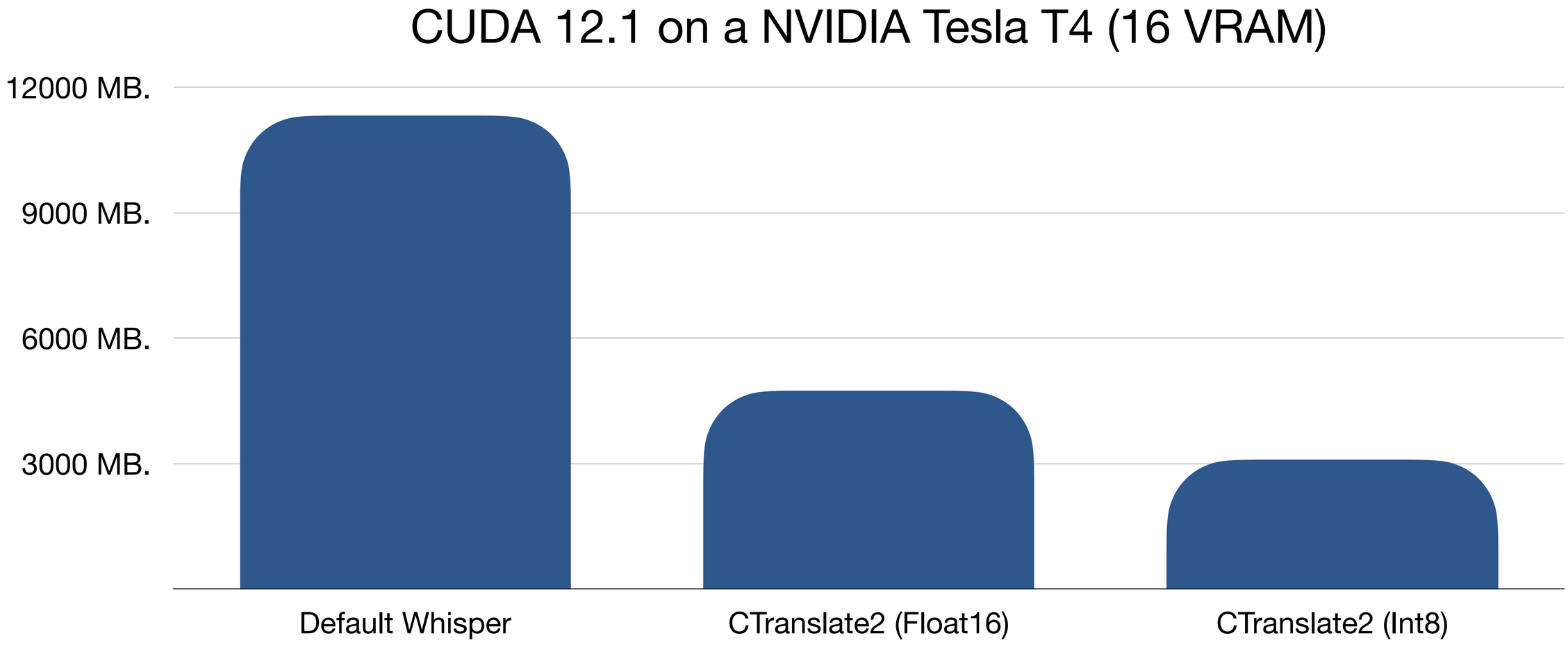
Работа с моделью



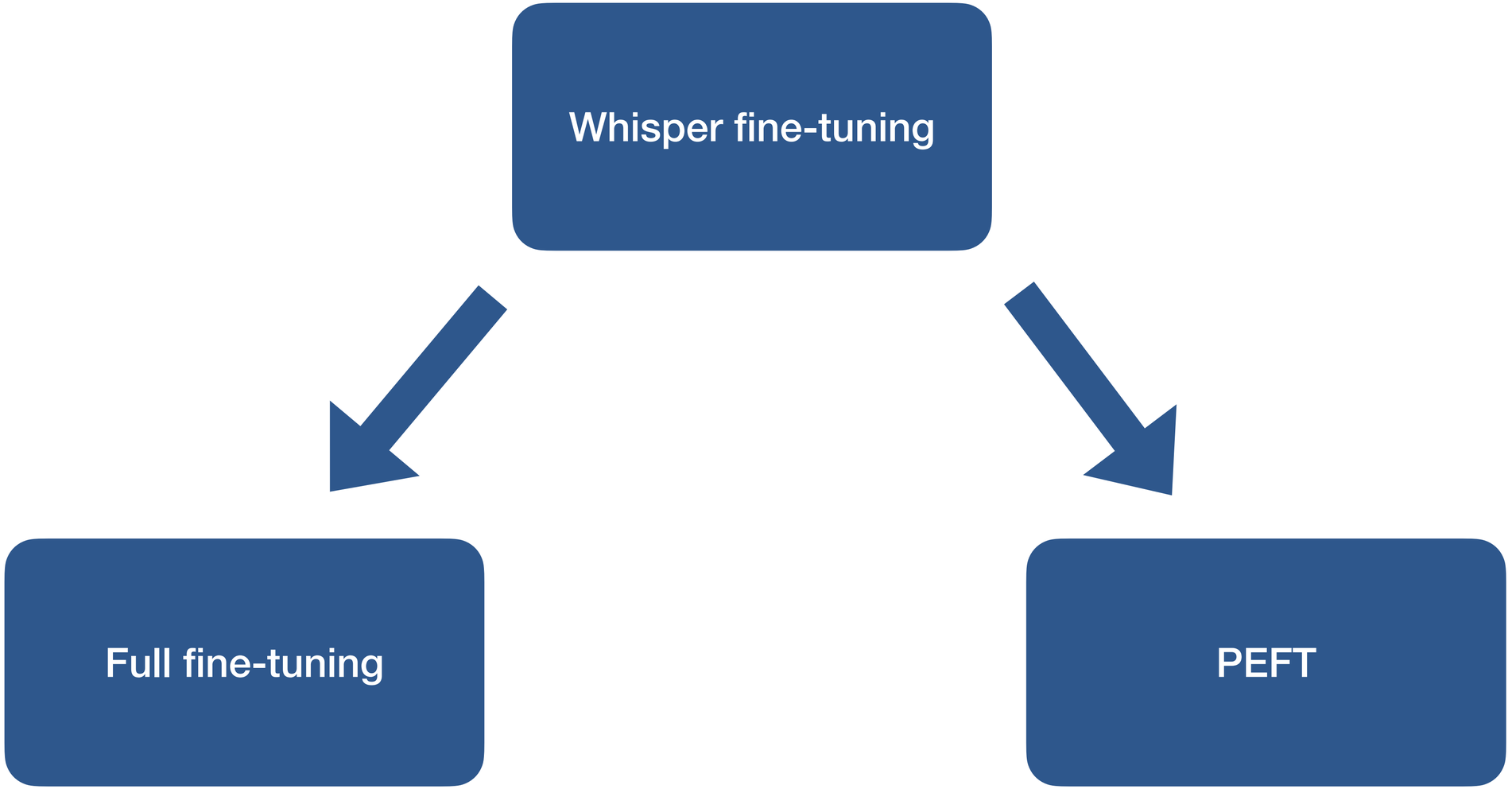
Оптимизация по скорости



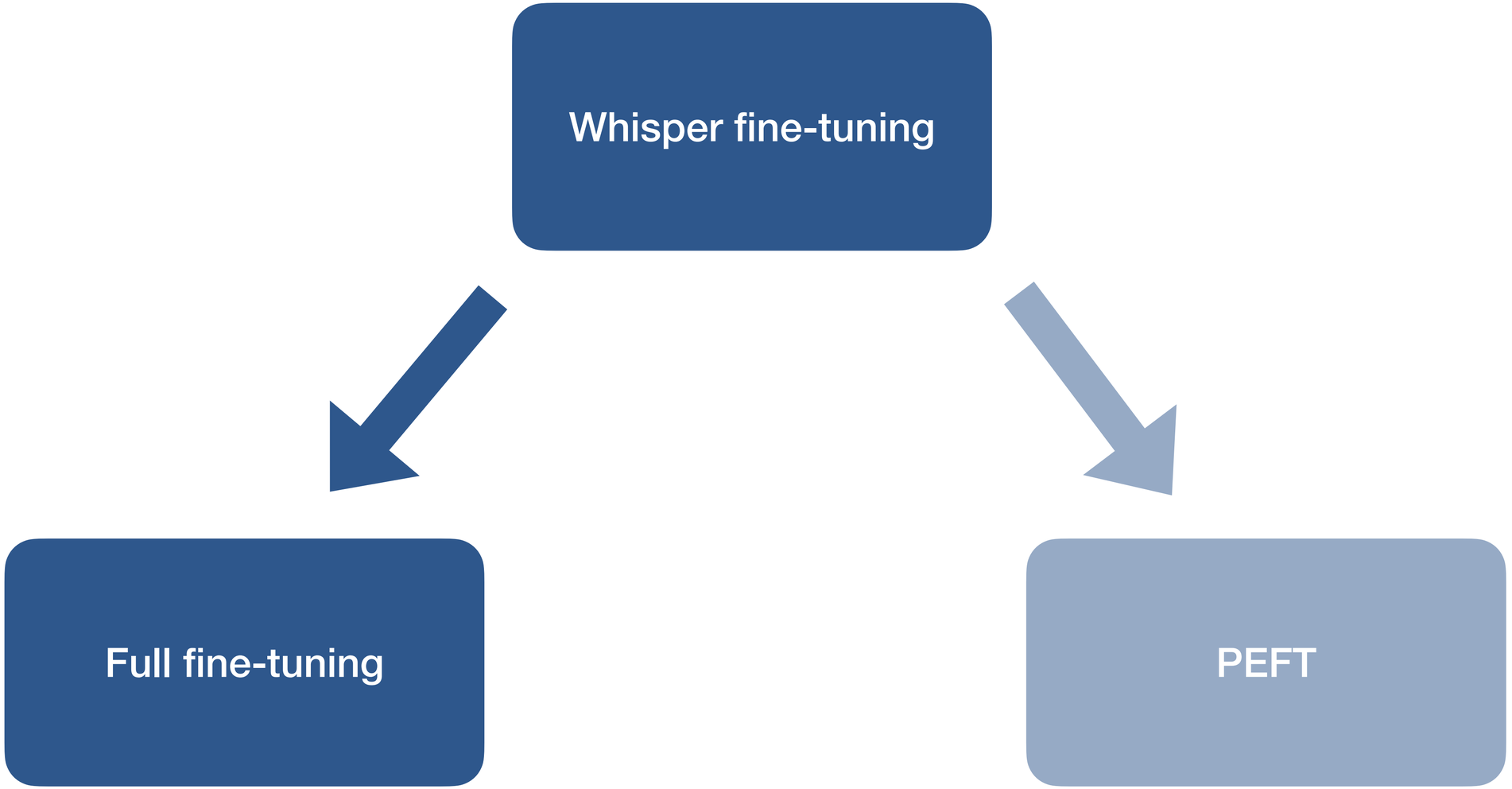
Оптимизация по ресурсам



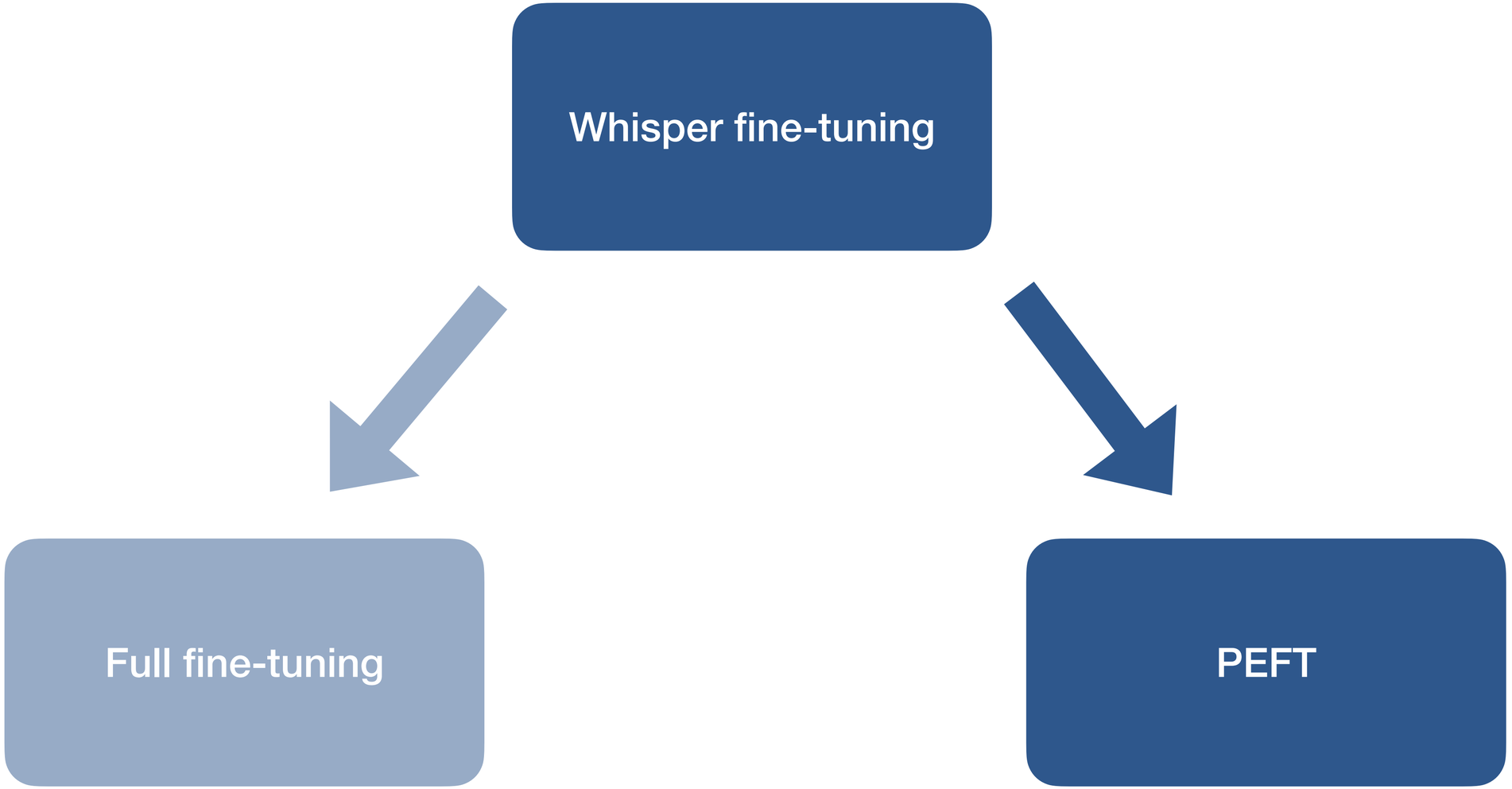
Вариации дообучения



Full fine-tuning

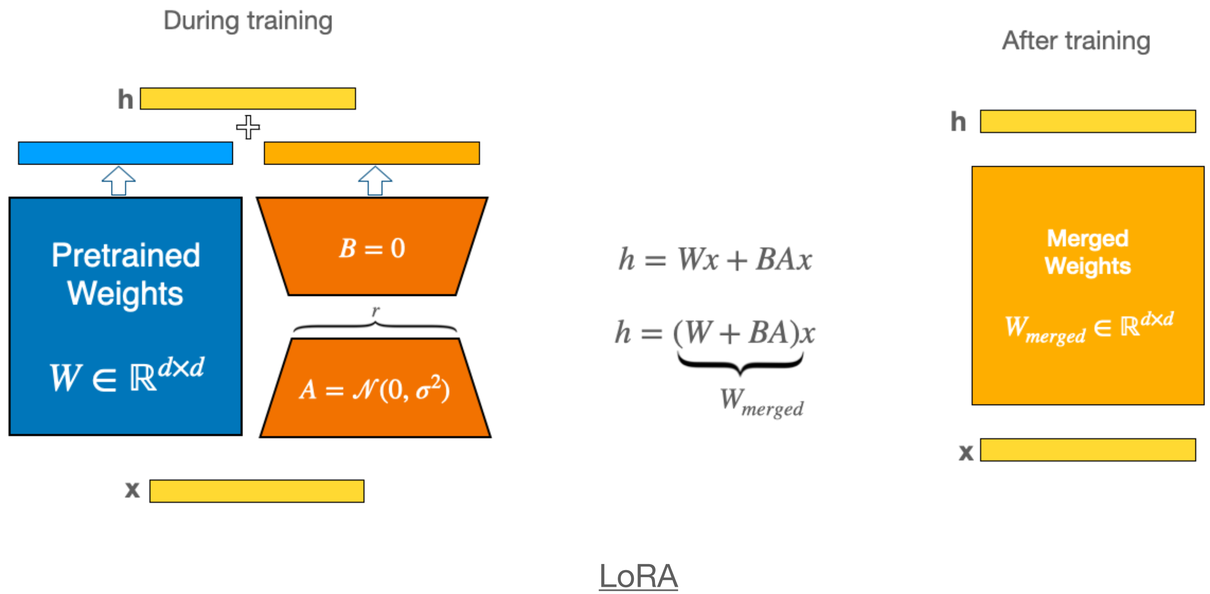


PEFT



LoRA powered by PEFT

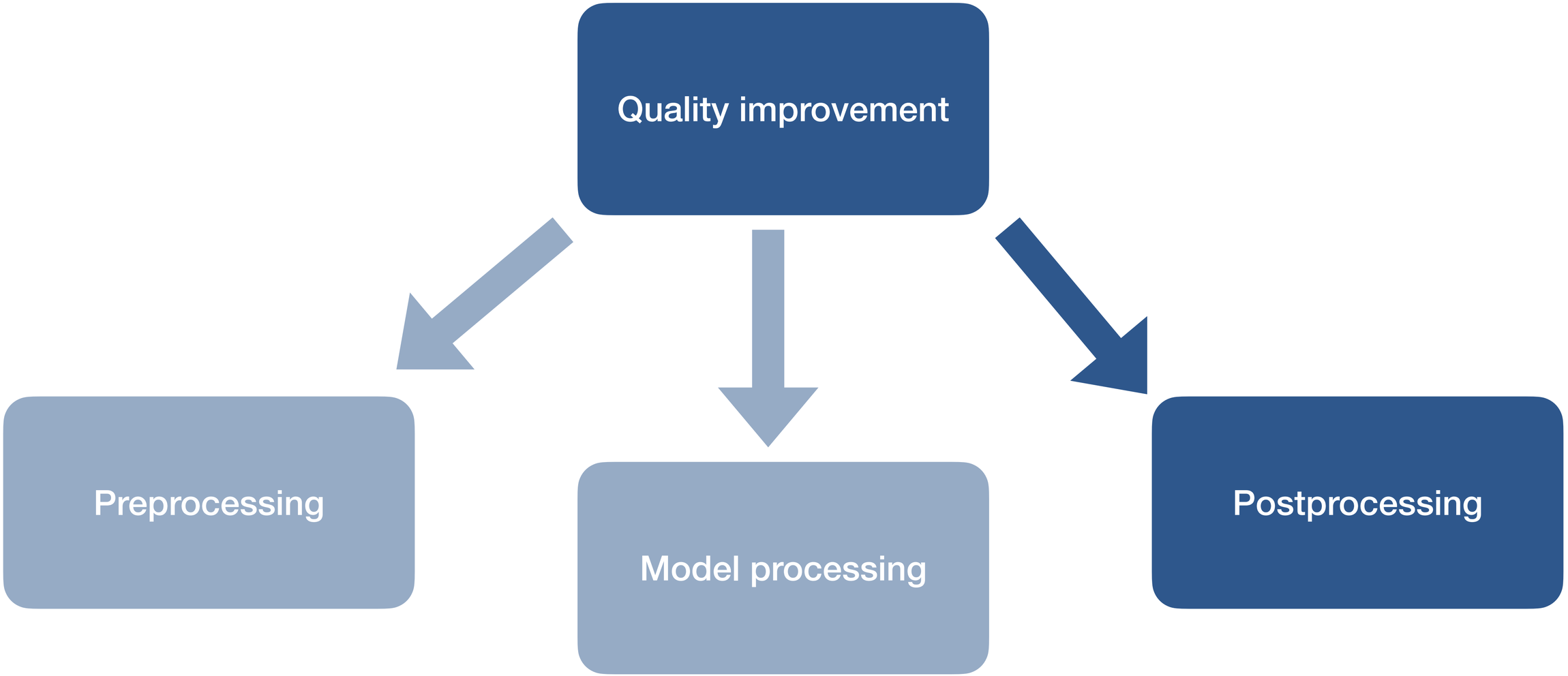
- Ускорение времени обучения
- Снижение затрат на вычисления и хранение
- Меньший риск переобучения
- Преодоление катастрофического забывания
- Удобство развертывания



LoRA

Fine-tune pipeline

Постпроцессинг



Alignment pipeline

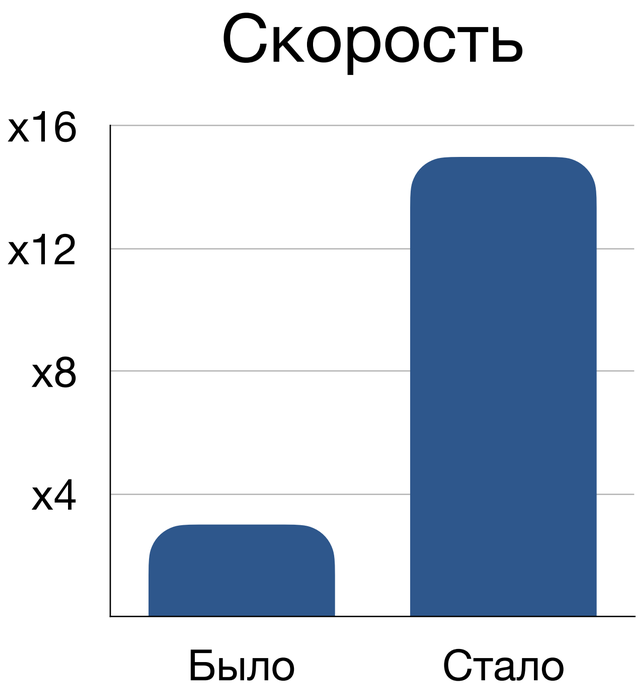
Good case #1

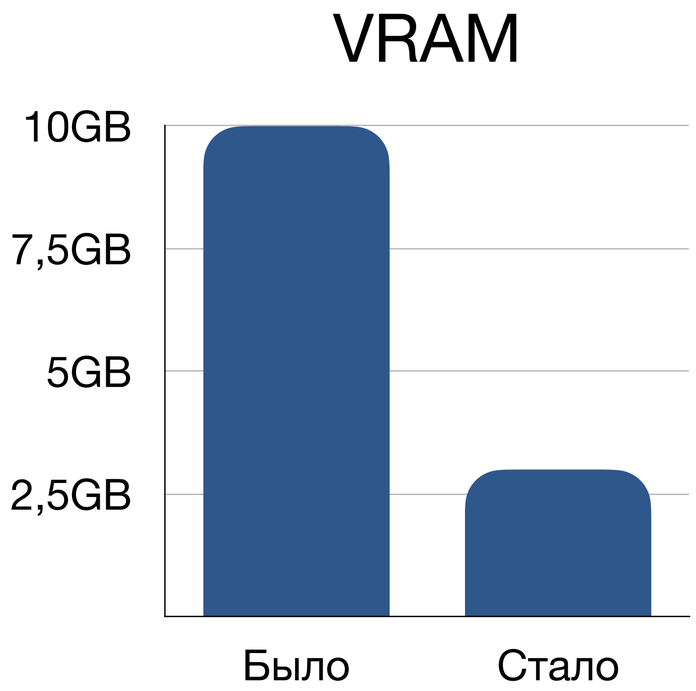
https://www.mosfilm.ru/

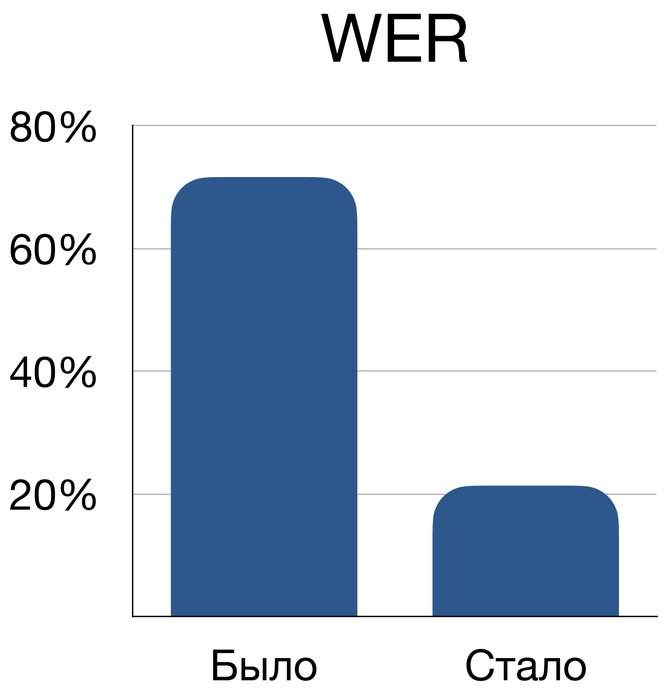
Good case #2

https://www.mosfilm.ru/

Итоги


Скорость
x16
x12
x8
x4
Было
Стало


VRAM
10GB
7,5GB
5GB
2,5GB
Было
Стало


WER
80%
60%
40%
20%
Было
Стало

Спасибо за внимание!

Мои контакты:

Почта: dmitry.shatnew@gmail.com