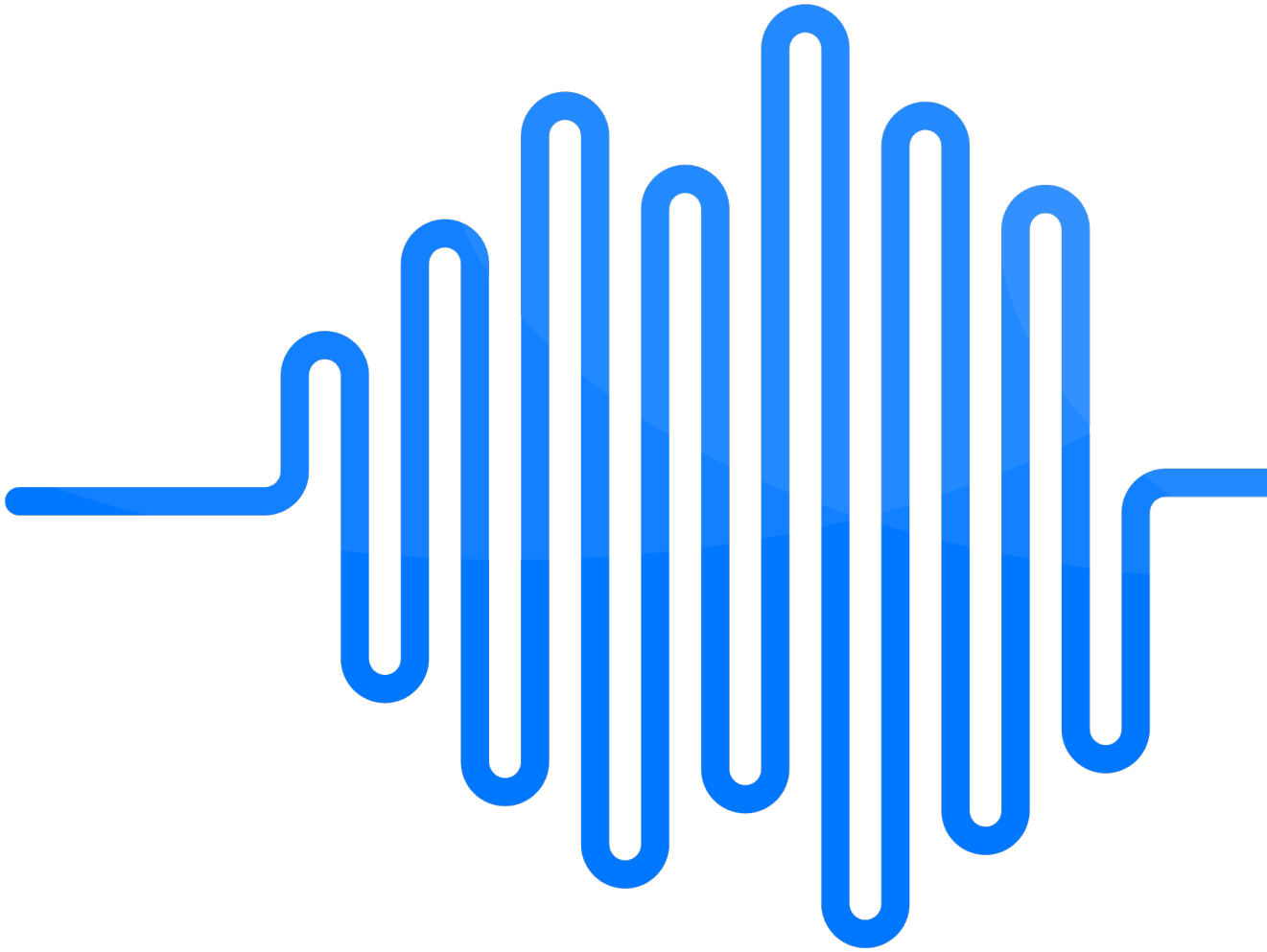
Метрика качества голоса ВК Звонков и как мы к ней пришли

Иван Бескровный,
ВКонтакте

Обо мне

- Инженер-разработчик в Команде звуковых технологий ВКонтакте
- Разрабатывал аудио алгоритмы для устройств Huawei
- 6+ опыта разработки в сфере аудио и речи
- Читаю курс по Audio ML в магистратуре МФТИ
- Сейчас занимаюсь обработкой голоса в ВК Звонках



Содержание

Для чего?

- Задача потокового аудио
- Слабые места алгоритмов обработки потокового аудио

1

Что это?

- Обзор оригинальной метрики
- Наши модификации
- Производительность на устройствах

2

Зачем?

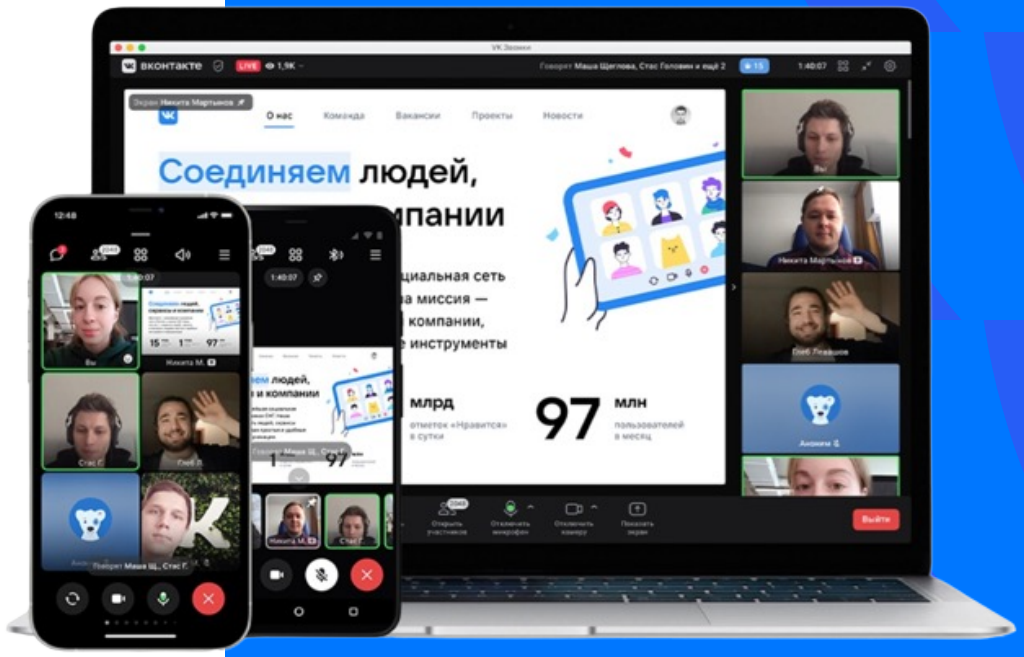
- Использование метрики: оценка каждого отдельного этапа обработки
- Сценарии использования: уведомления пользователю, трейсинг событий в поддержке

3

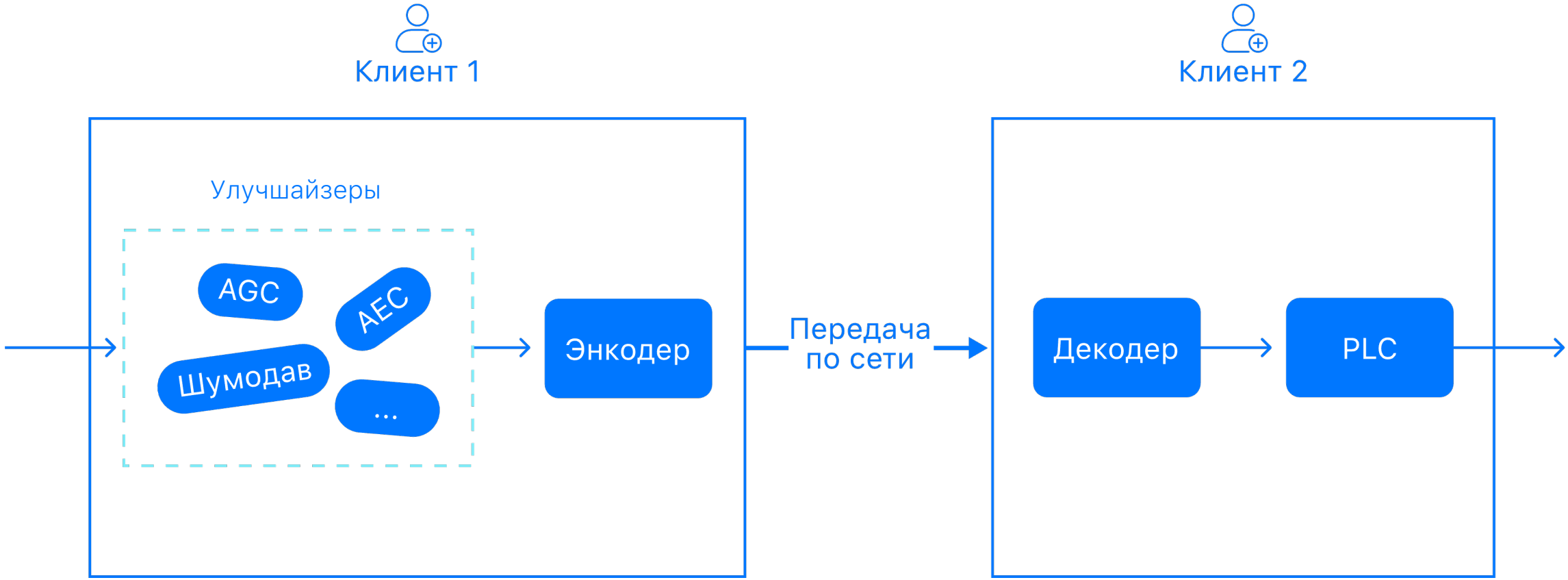
Для чего?

Задача потокового аудио

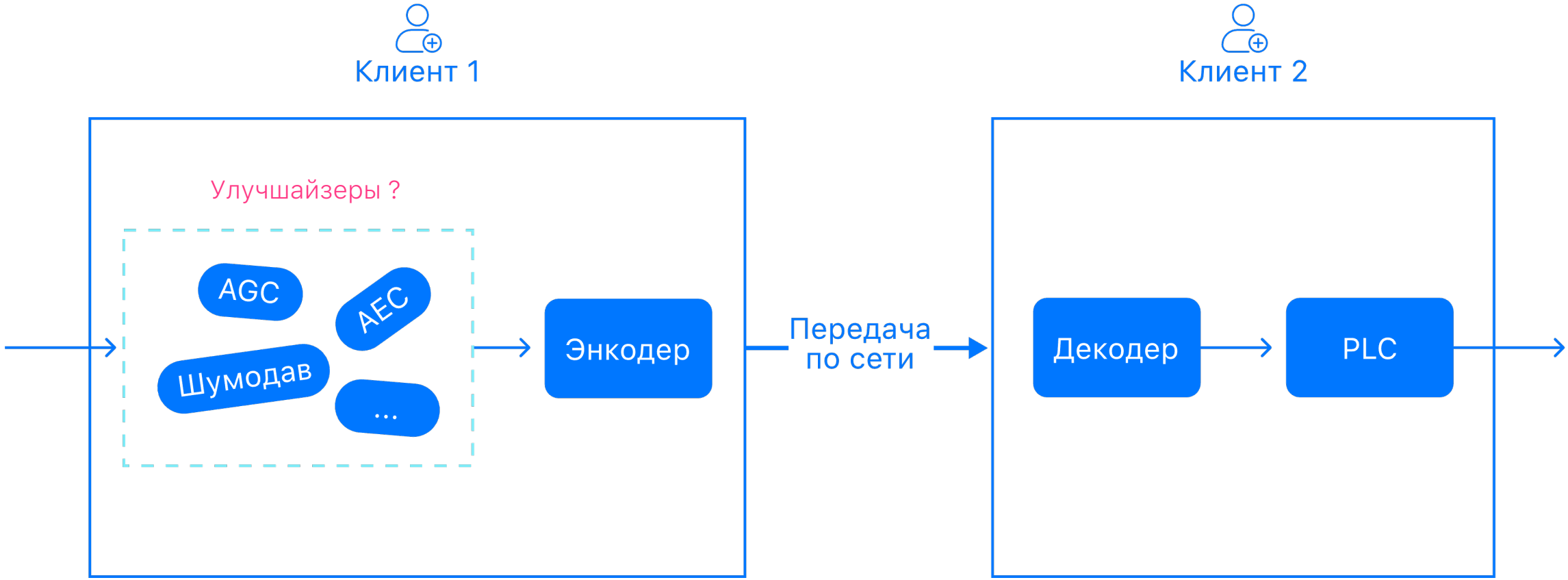
- Звонки
- Трансляции
- Голосовые сообщения



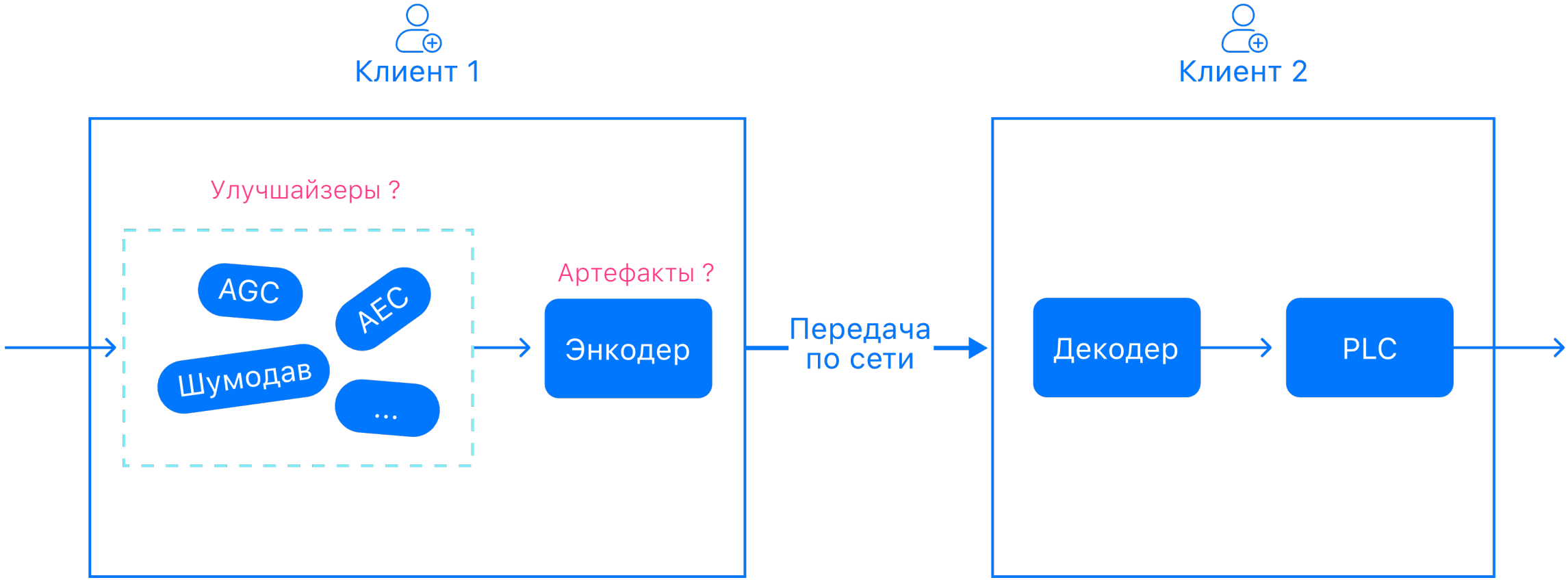
Пайплайны обработки потокового аудио



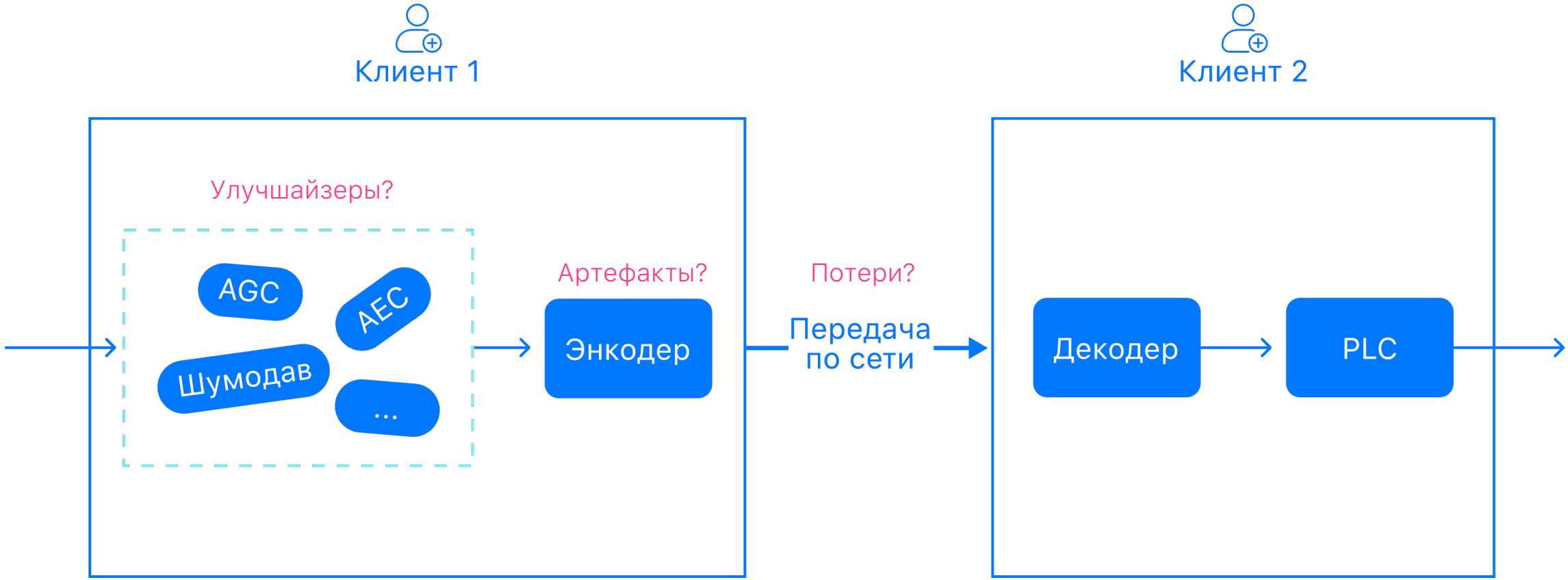
Пайплайны обработки потокового аудио



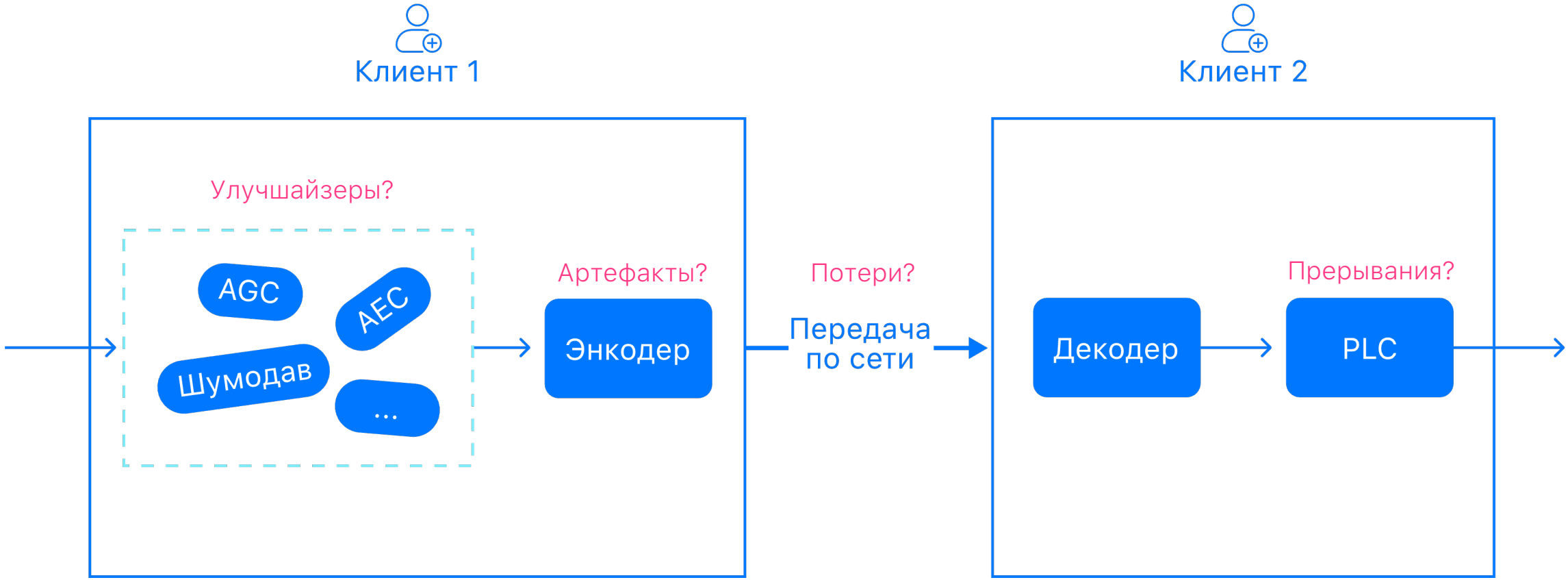
Пайплайны обработки потокового аудио



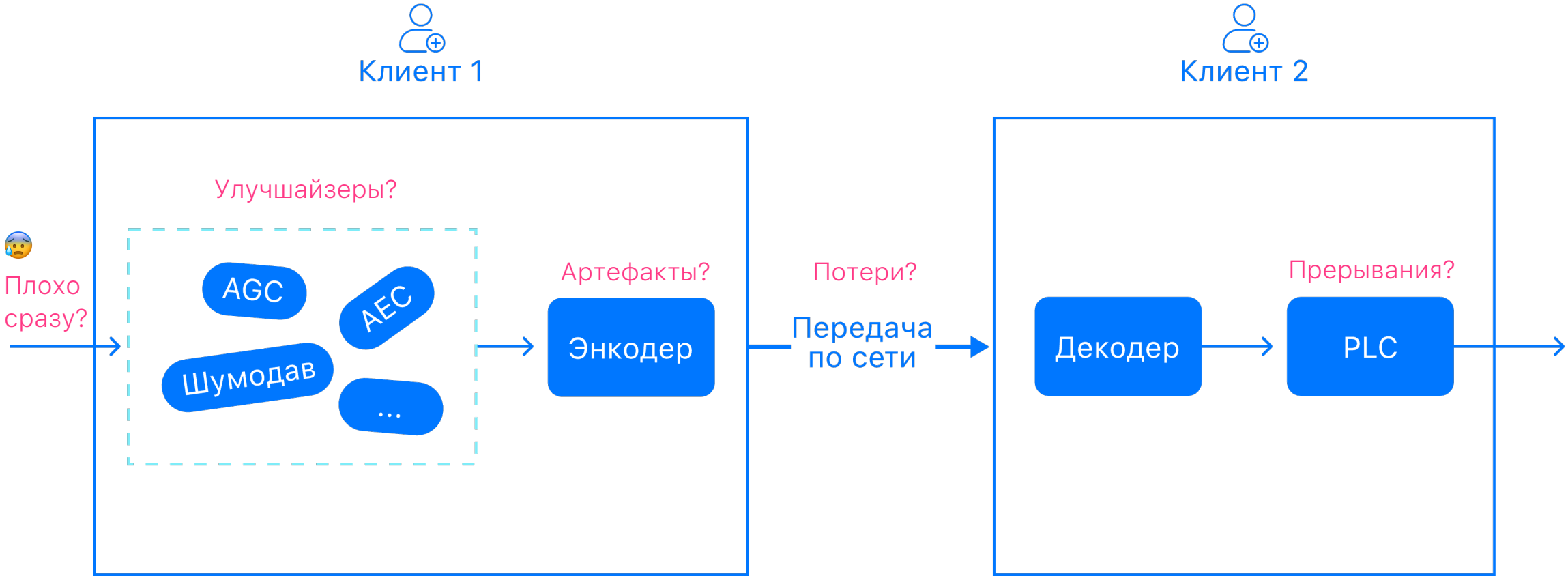
Пайплайны обработки потокового аудио



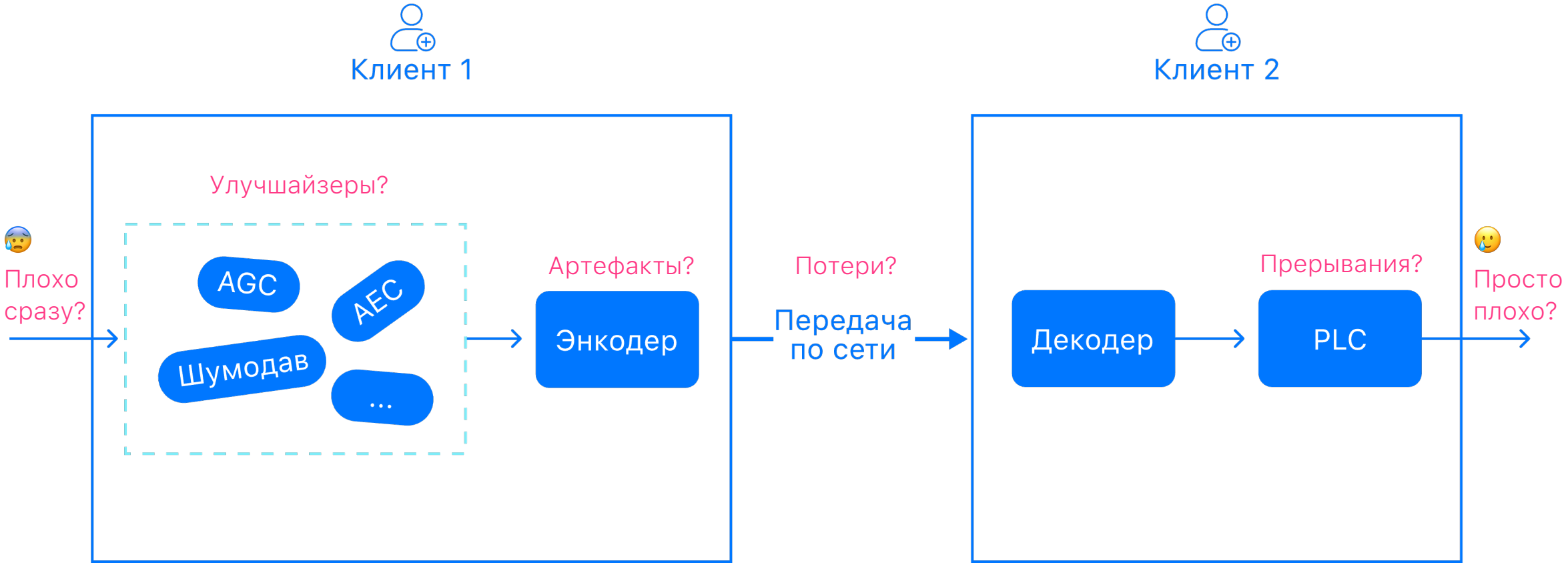
Пайплайны обработки потокового аудио



Пайплайны обработки потокового аудио

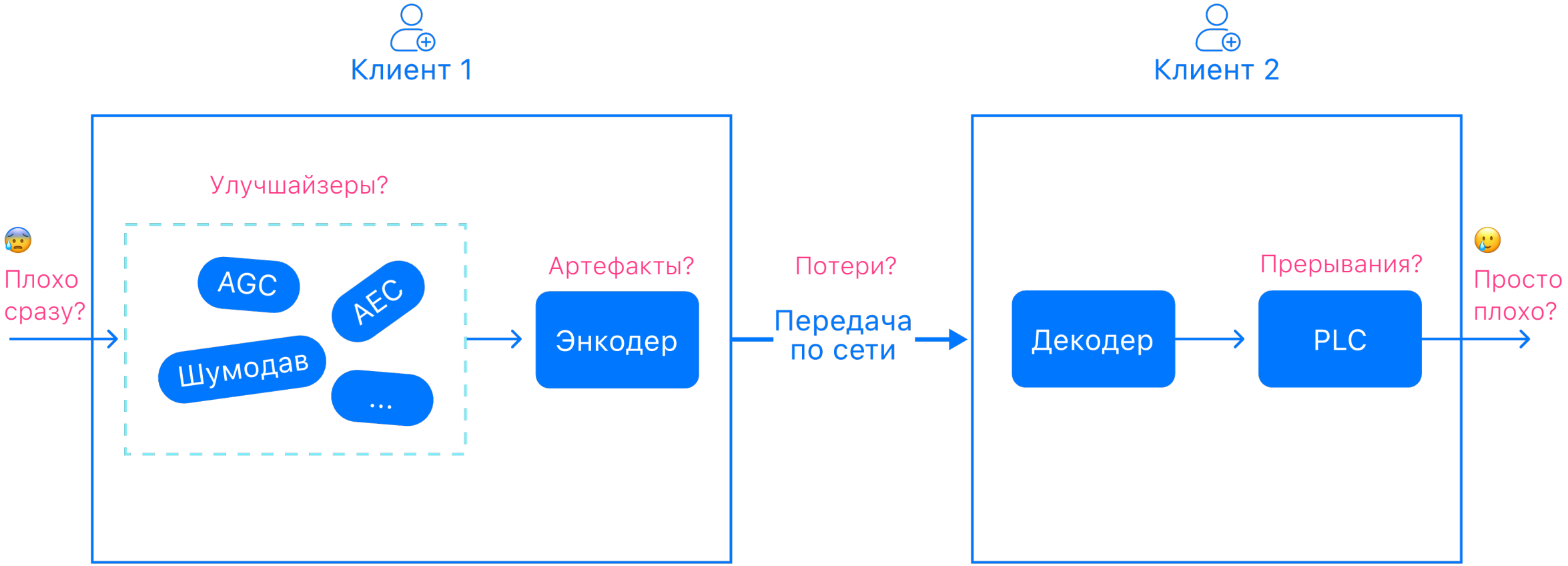


Пайплайны обработки потокового аудио



Пайплайны обработки потокового аудио

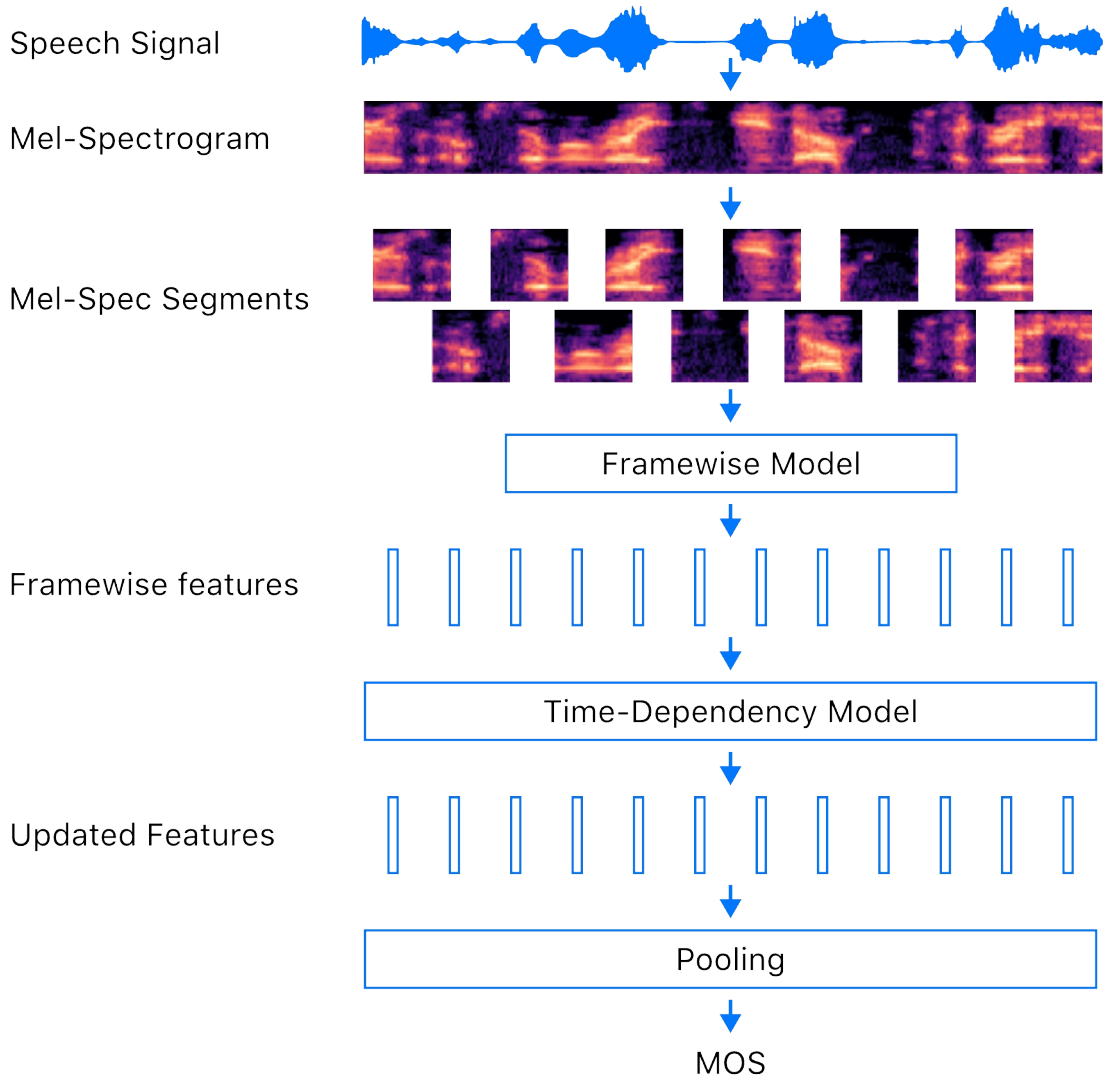
Вывод: хочется как-то контролировать качество на каждом этапе



Что это?

Оригинальная метрика NISQA

- https://arxiv.org/pdf/2104.09494
- Три вида сверточных блоков
- Attention в качестве временных блоков
- Пять метрик на выходе:
  - Noisiness
  - Coloration
  - Discontinuity
  - Loudness
  - MOS



Почему не взяли «из коробки»?

1. Низкая скорость работы

2. Слабо читаемый код
   - Два файла на 1000+ и 2000+ строк кода

3. Лишние блоки
   - Свертки с дублирующимся функционалом
   - Аттеншен вместе с LSTM

4. Хардкод параметров
   - Фиксированное окно STFT
   - Фиксированная длина семпла на инференсе
   - Фиксированный размер сверток

Почему не взяли «из коробки»?

1. Низкая скорость работы

2. Слабо читаемый код
   - Два файла на 1000+ и 2000+ строк кода

3. Лишние блоки
   - Свертки с дублирующимся функционалом
   - Аттеншен вместе с LSTM

4. Хардкод параметров
   - Фиксированное окно STFT
   - Фиксированная длина семпла на инференсе
   - Фиксированный размер сверток

Что мы поменяли?

1. Добавили «онлайн»
   - Интерфейс фреймами от 20 мс
   - Захват сигнала с системного микрофона

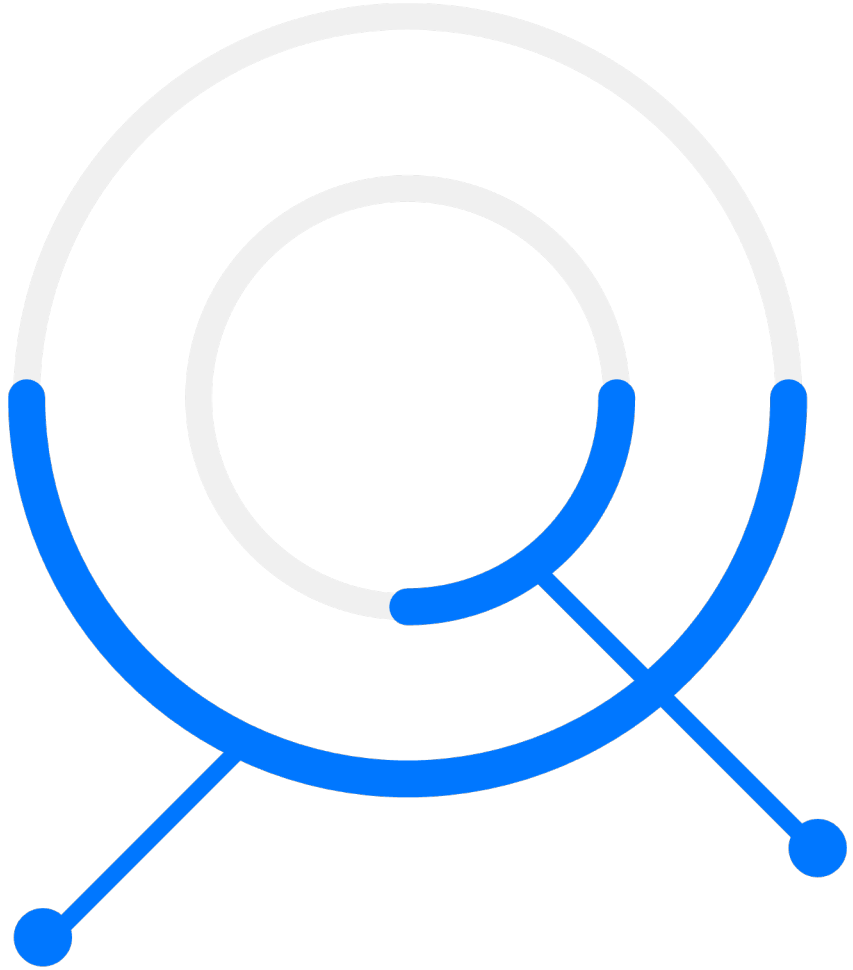
2. Избавились от дублирующихся блоков
   - CNN / AdaptCNN / SkipCNN -> AdaptCNN
   - Attention / LSTM -> LSTM

3. Облегчили оставшиеся
   - Меньше свертки
   - Меньше окно STF

4. Модифицировали временные блоки
   - Стейты LSTM

Что получили?

- Значения метрик совпадают
  с коробочной версией
  до второго знака после запятой
- 40-70 реалтаймов*
  (по сравнению с 5-10
  у коробочной версии)
- Инференс сигнала
  напрямую с микрофона
- Читаемый код :)



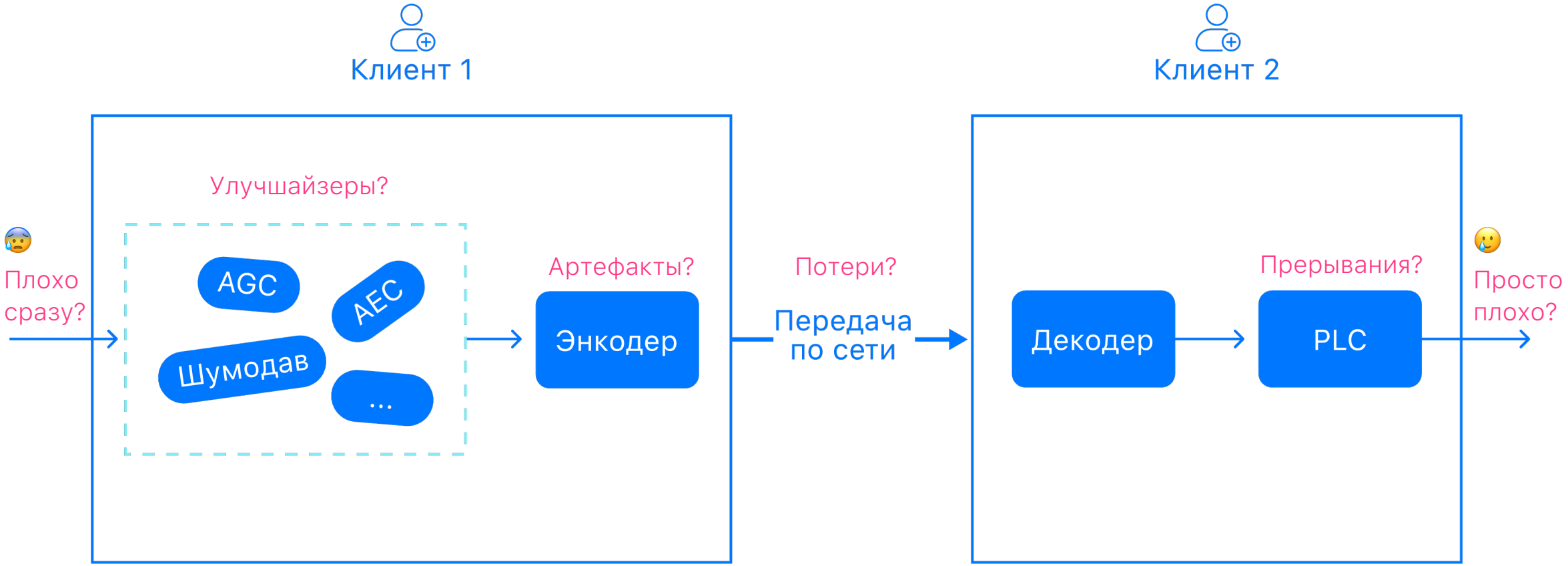
*на M1

Зачем?

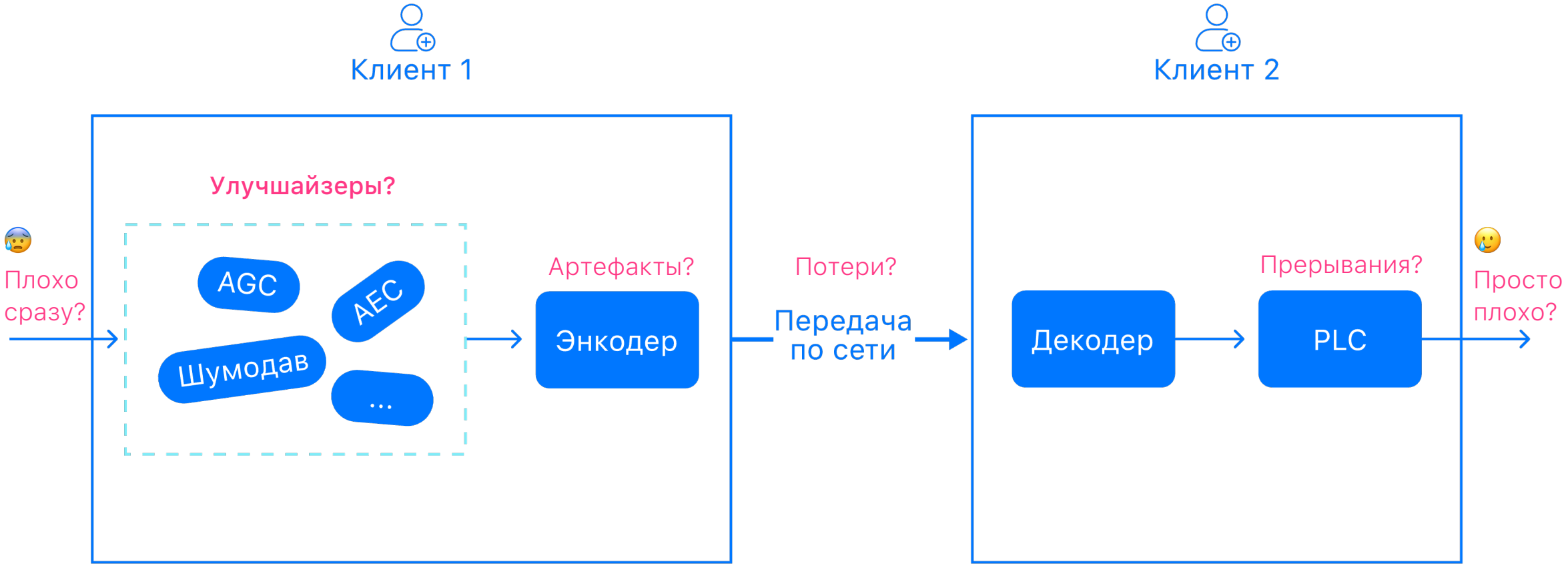
Плохо бывает разное

- Слишком много шума
- Слишком тихо
- Звук «как из бочки»
- Звук прерывается
- Просто «плохо»

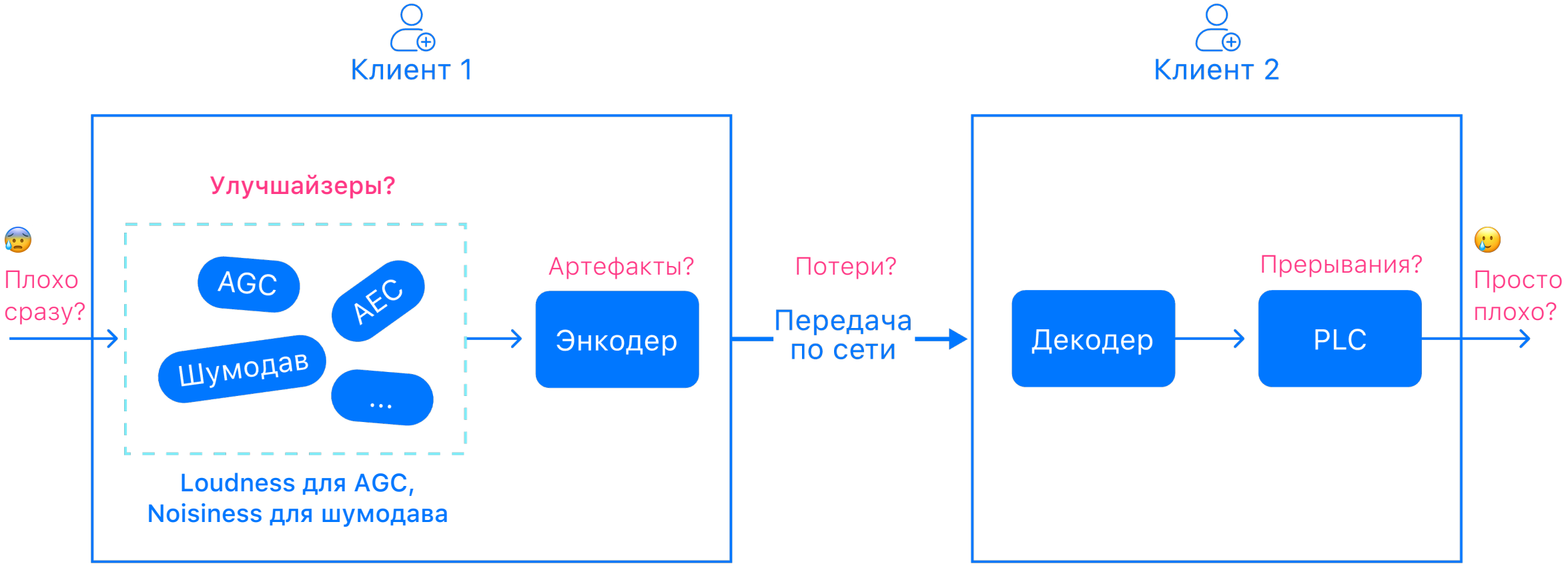
Каждый из этапов пайплайна можно проверить:



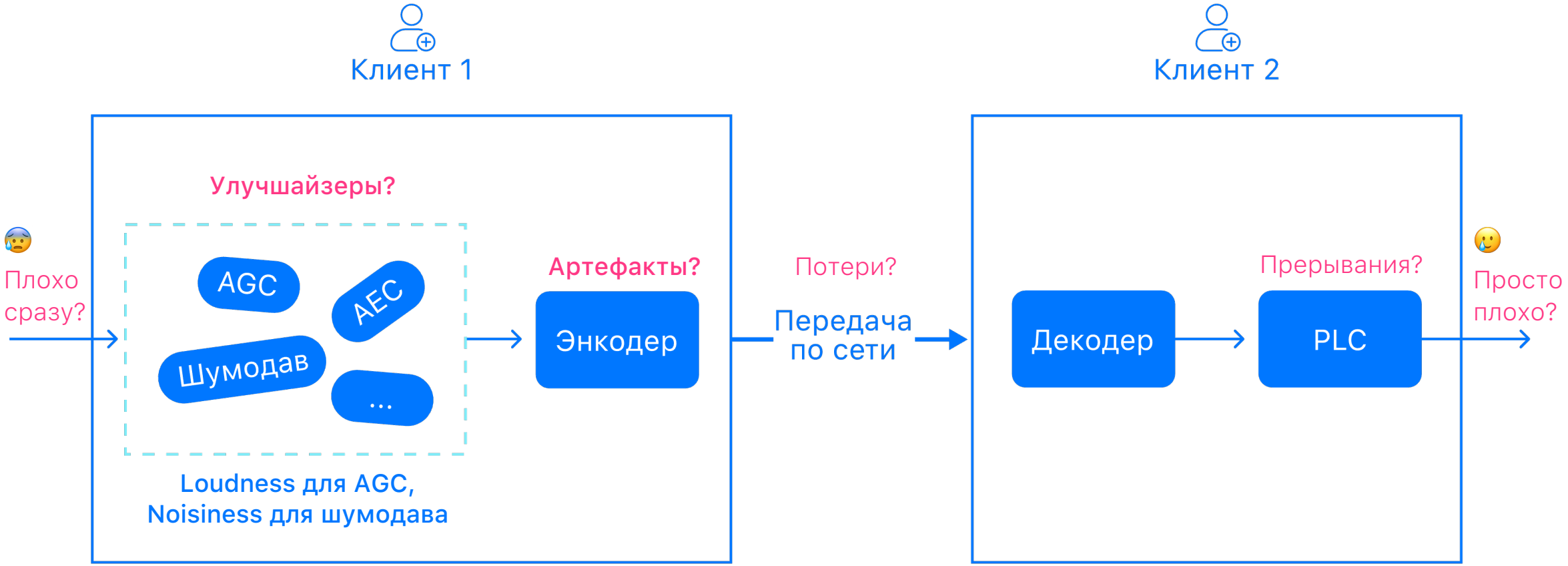
Каждый из этапов пайплайна можно проверить:



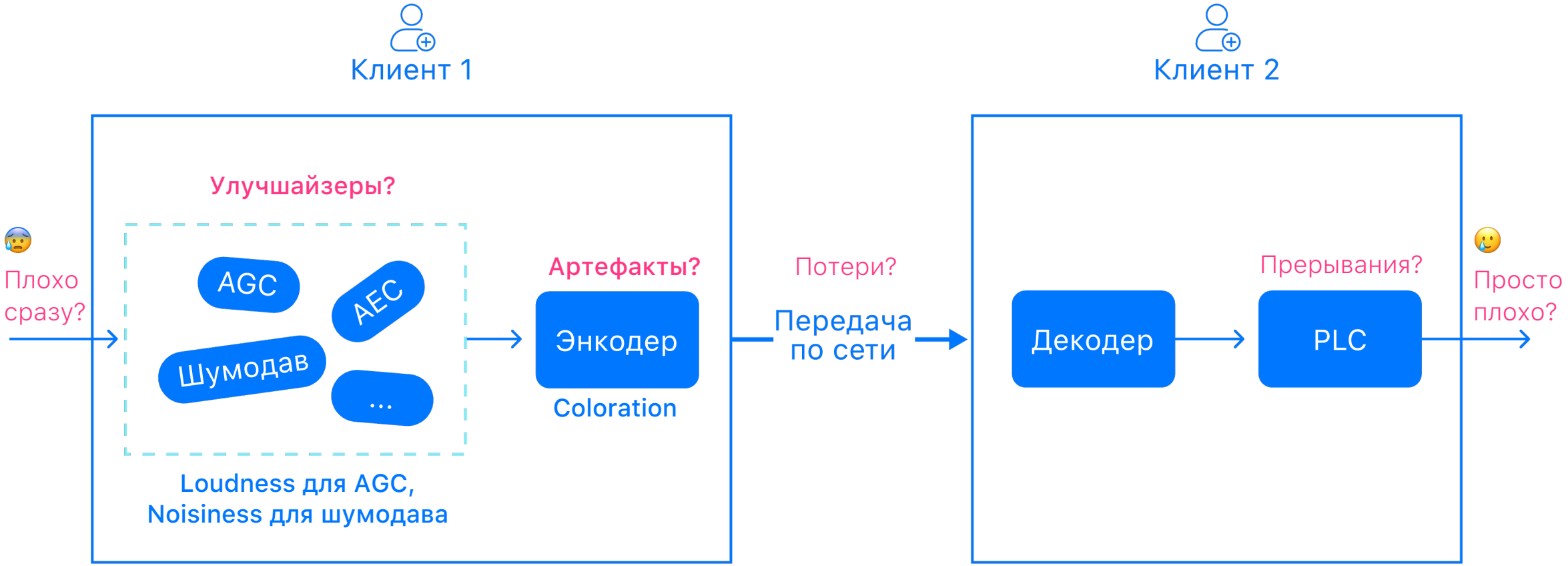
Каждый из этапов пайплайна можно проверить:



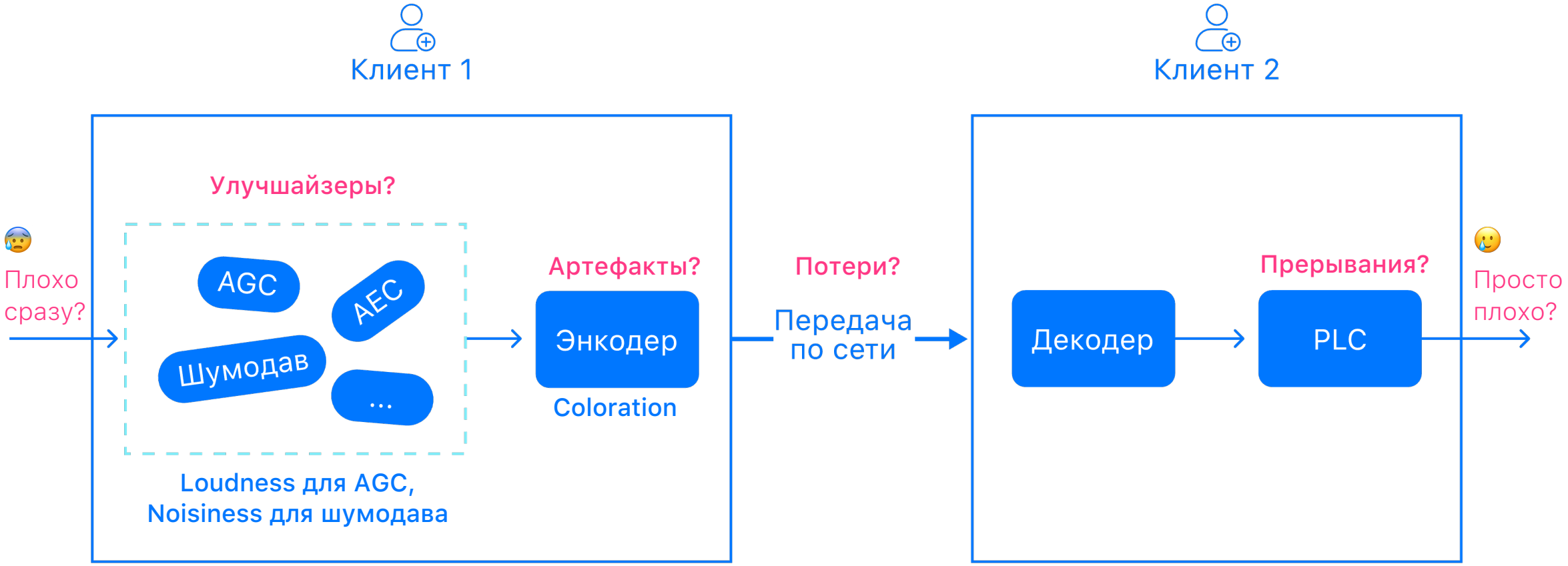
Каждый из этапов пайплайна можно проверить:



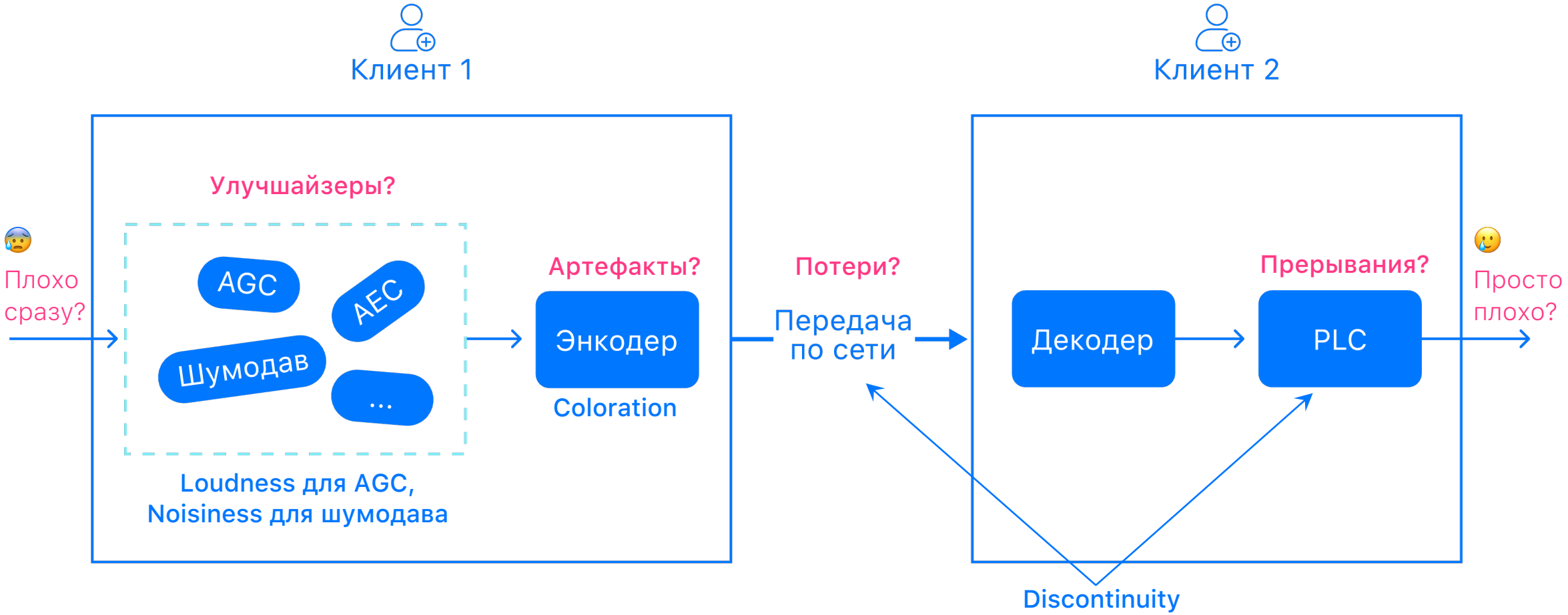
Каждый из этапов пайплайна можно проверить:



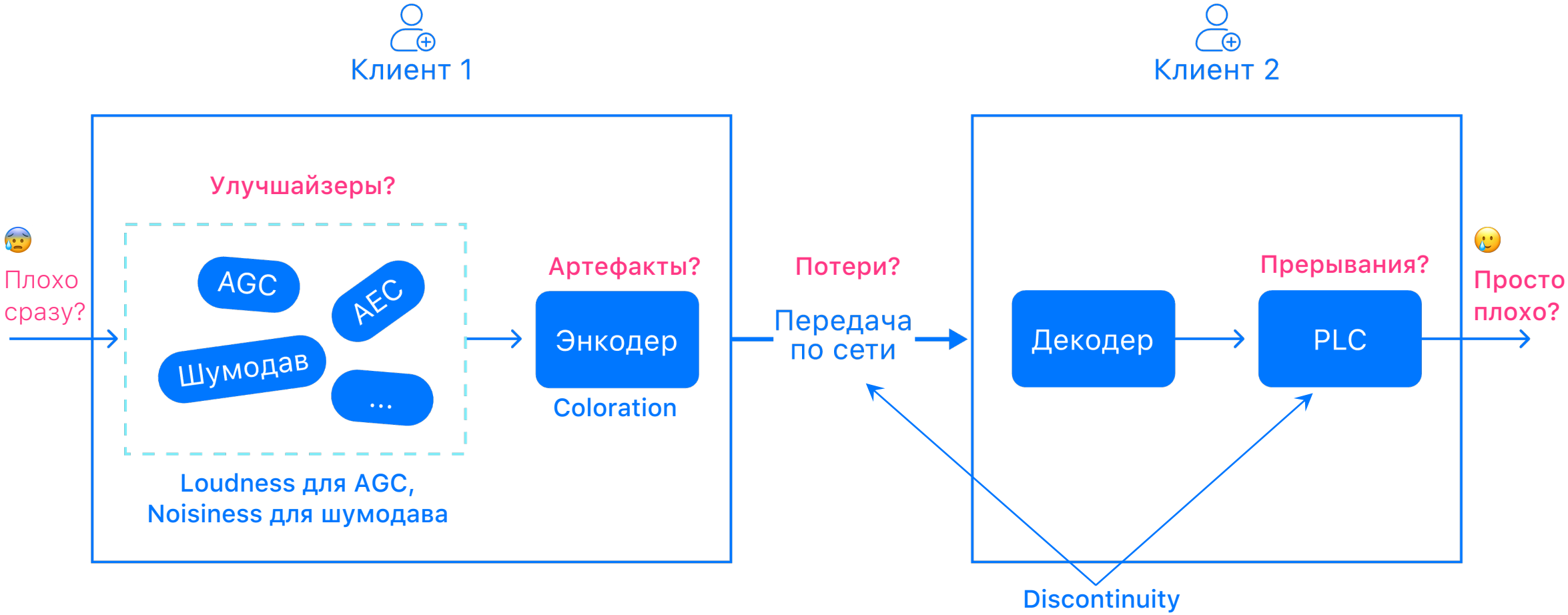
Каждый из этапов пайплайна можно проверить:



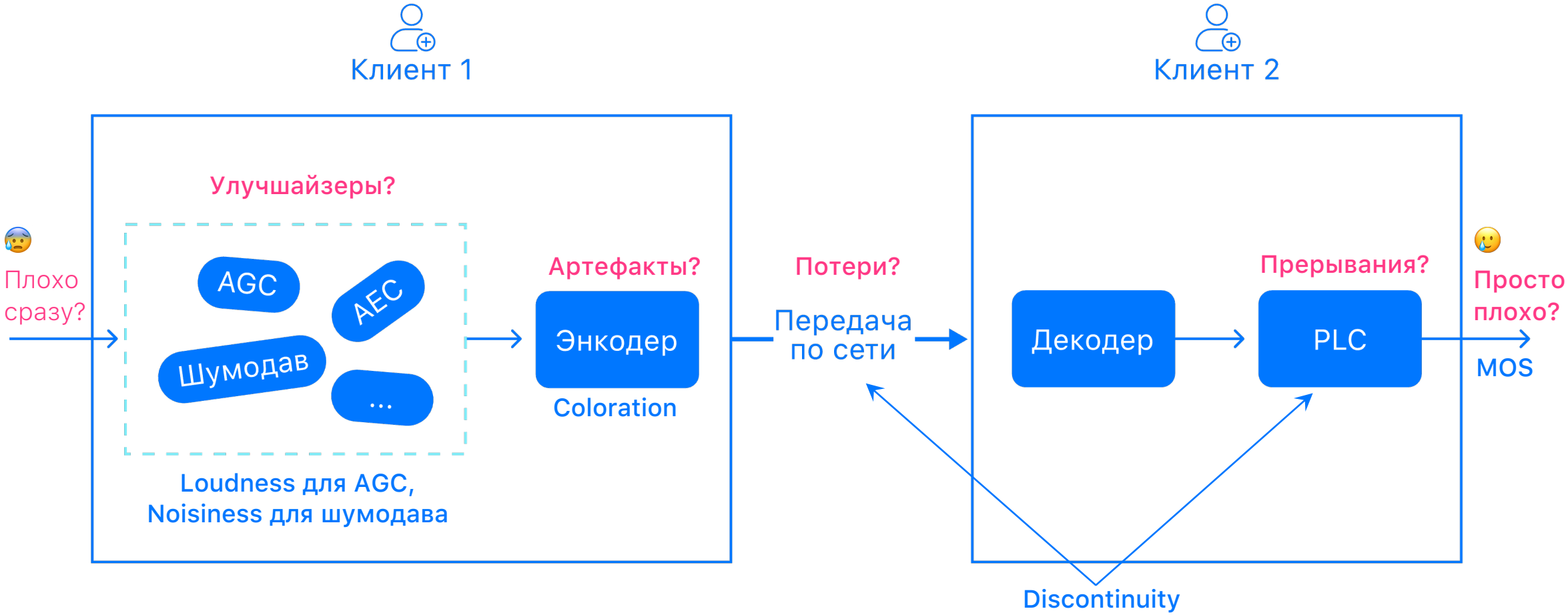
Каждый из этапов пайплайна можно проверить:



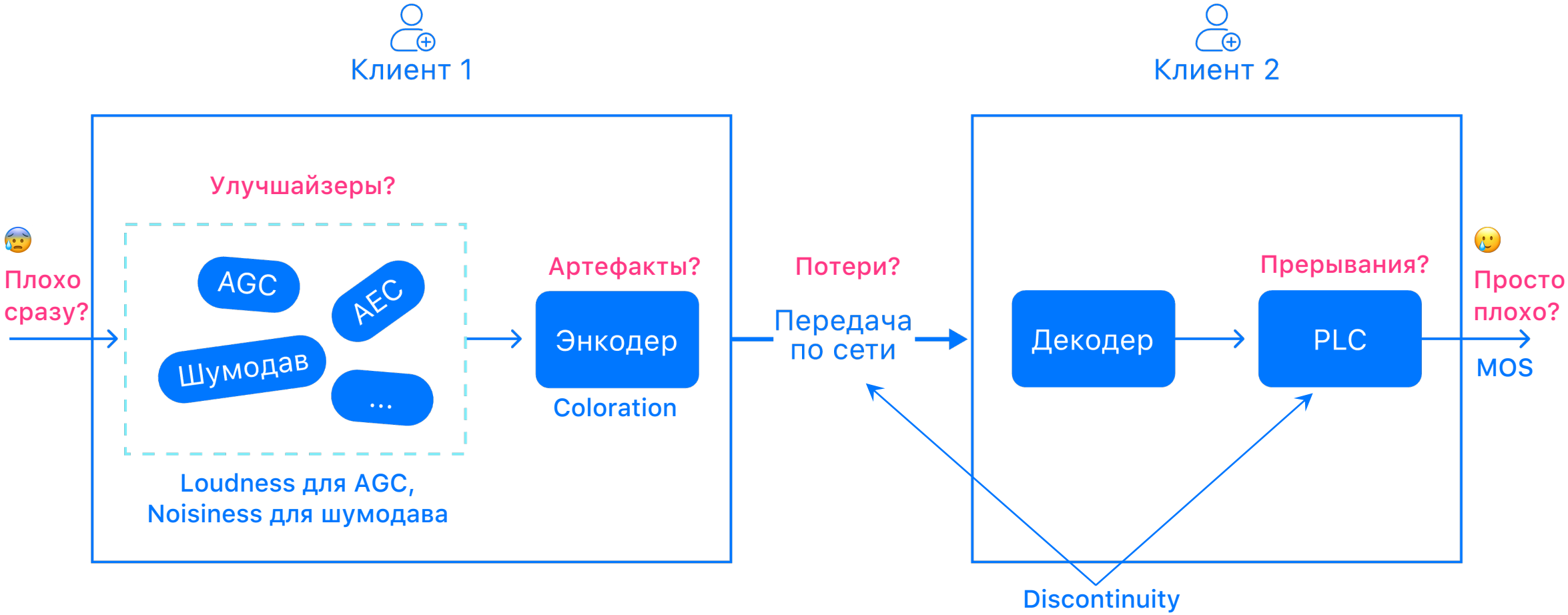
Каждый из этапов пайплайна можно проверить:



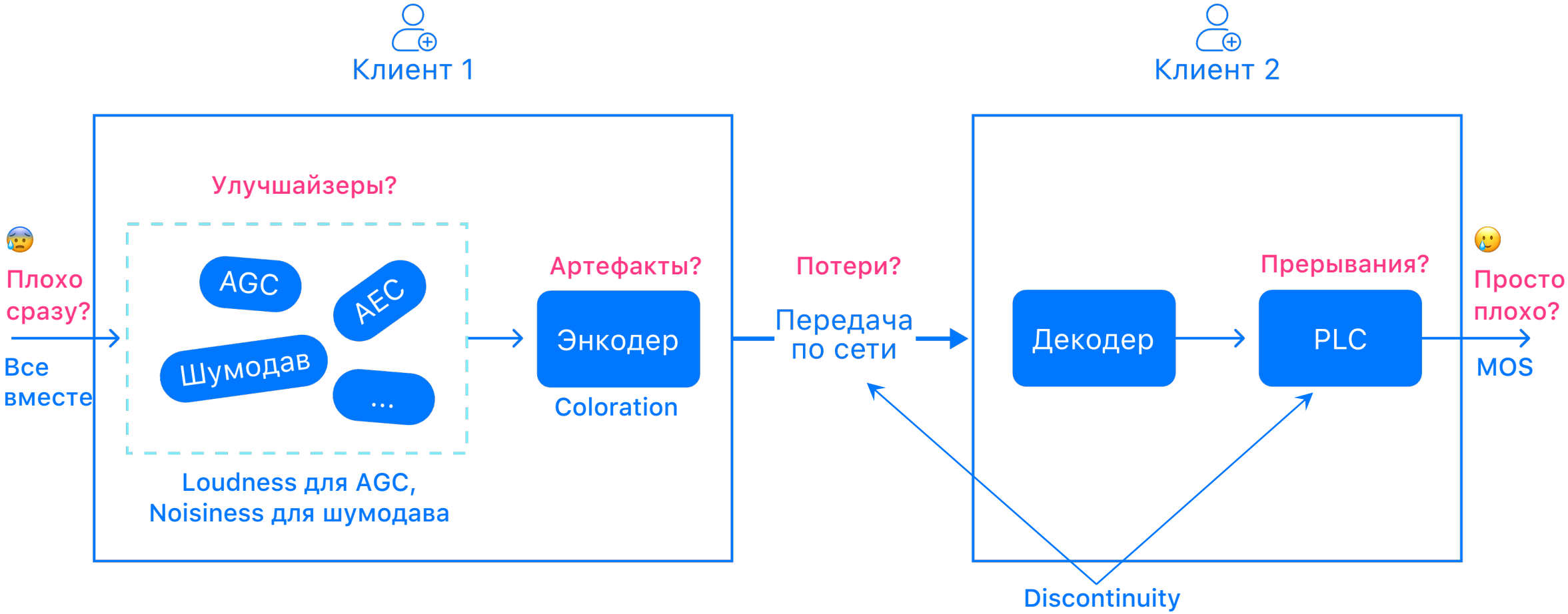
Каждый из этапов пайплайна можно проверить:



Каждый из этапов пайплайна можно проверить:

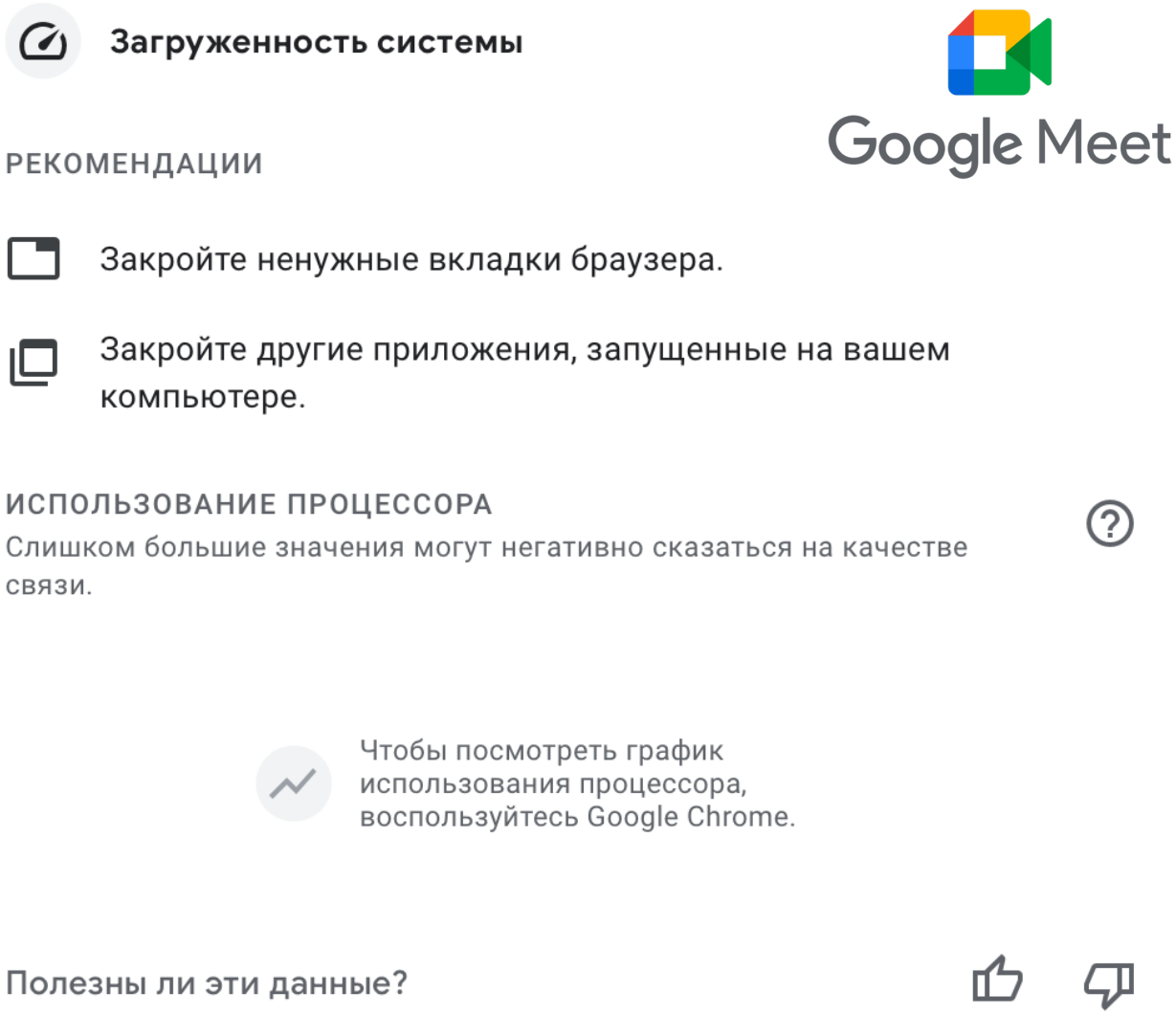


Каждый из этапов пайплайна можно проверить:



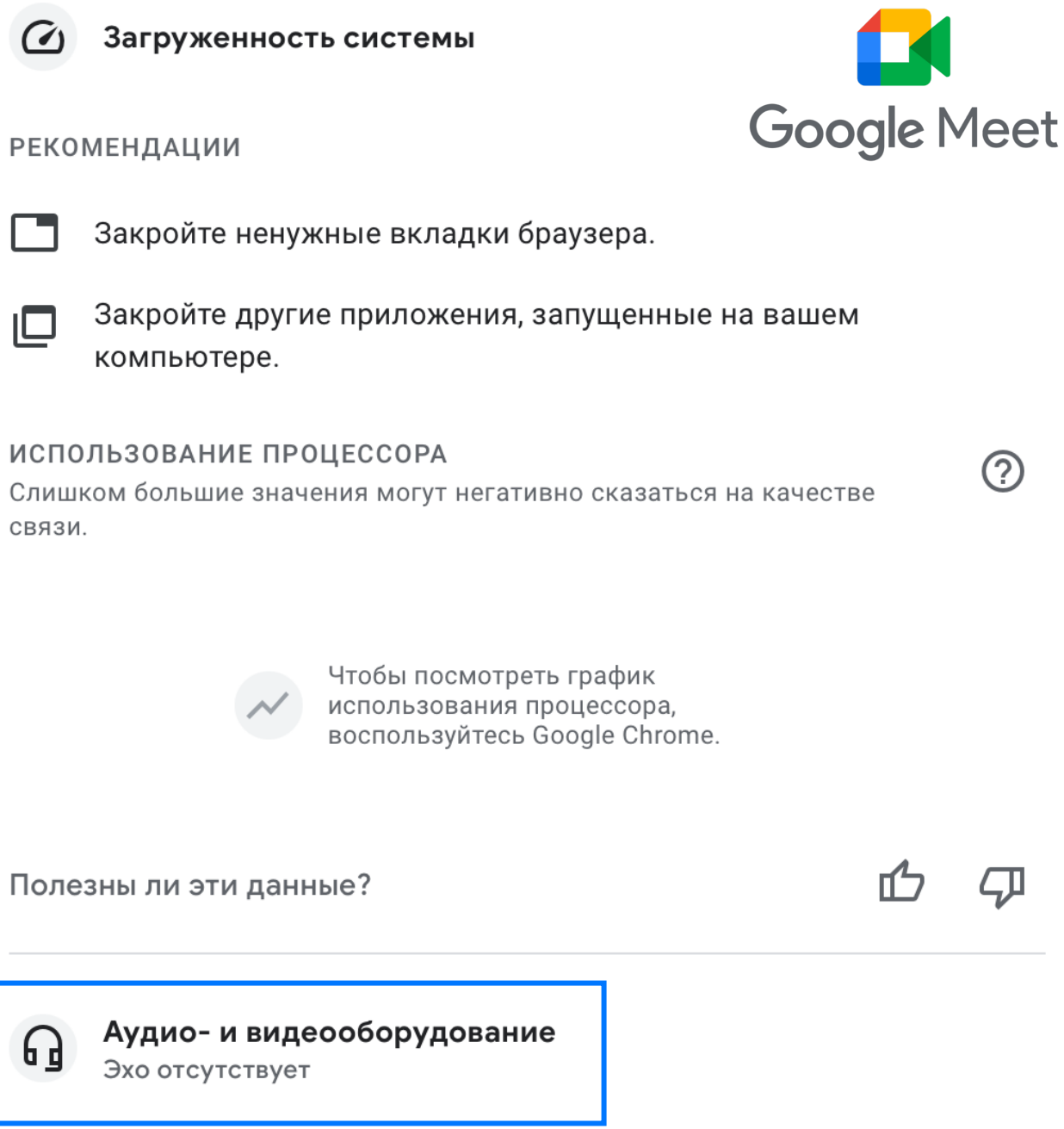
Сценарии использования

Статистика и / или уведомления пользователю («плохо сразу»)



Сценарии использования

Статистика и / или уведомления пользователю («плохо сразу»)



Сценарии использования

Системы трейсинга событий («плохо» где-то внутри)

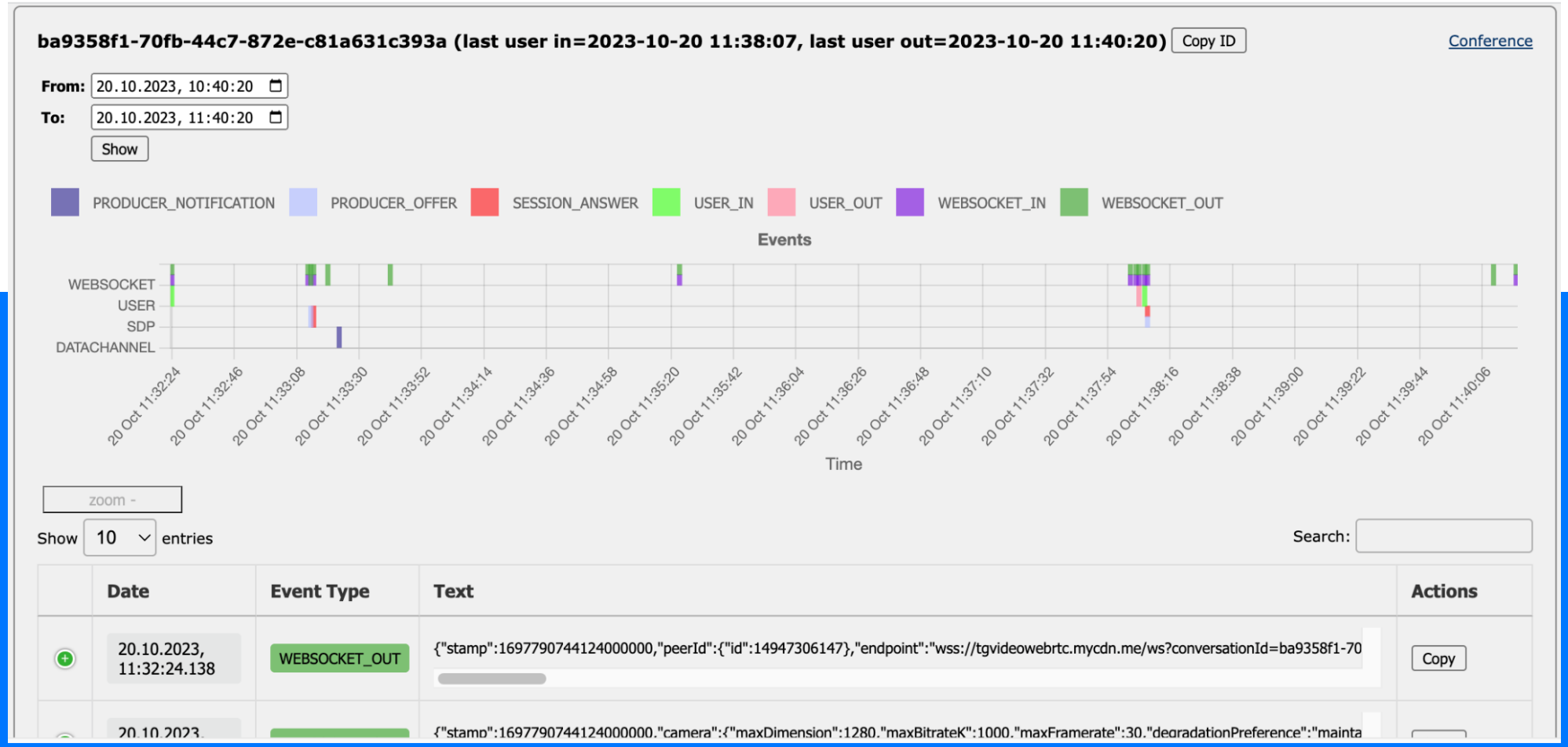
Палантир: короткое интро

Что за Палантир?

- Система трейсинга событий в бэкенде ВК Звонков
- Помогает поддержке отслеживать «плохо» в звонках
- Логгирует события в каждом отдельном звонке

USER_IN	пользователь вошел в звонок
USER_OUT	пользователь вышел из звонка
WEBSOCKET_IN	сообщение по WS от клиента к серверу
WEBSOCKET_OUT	сообщение по WS от сервера к клиенту
PRODUCER_OFFER	SDP offer к клиенту
SESSION_ANSWER	SDP answer от клиента

Что за Палантир?



NISQA-s + Палантир = 💙

- В текущей реализации нет мониторинга качества
- Обращения пользователей крайне расплывчаты:

NISQA-s + Палантир = 💙

- В текущей реализации нет мониторинга качества
- Обращения пользователей крайне расплывчаты:

«Ваню Иванова было слышно через слово»

NISQA-s + Палантир = 💙

- В текущей реализации нет мониторинга качества
- Обращения пользователей крайне расплывчаты:

«Ваню Иванова было слышно через слово»

Смотрим на Discontinuity, проверяем PLC и сеть

NISQA-s + Палантир = 💙

- В текущей реализации нет мониторинга качества
- Обращения пользователей крайне расплывчаты:

«Ваню Иванова было слышно через слово»

«Света Светина была слишком тихая»

Смотрим на Discontinuity, проверяем PLC и сеть

NISQA-s + Палантир = 💙

- В текущей реализации нет мониторинга качества
- Обращения пользователей крайне расплывчаты:

«Ваню Иванова было слышно через слово»

«Света Светина была слишком тихая»

Смотрим на Discontinuity, проверяем PLC и сеть

Смотри на Loudness, проверяем параметры AGC

NISQA-s + Палантир = 💙

- В текущей реализации нет мониторинга качества
- Обращения пользователей крайне расплывчаты:

«Ваню Иванова было слышно через слово»

«Света Светина была слишком тихая»

«У Коли Николаева звук был как из бочки»

Смотрим на Discontinuity, проверяем PLC и сеть

Смотри на Loudness, проверяем параметры AGC

NISQA-s + Палантир = 💙

- В текущей реализации нет мониторинга качества
- Обращения пользователей крайне расплывчаты:

«Ваню Иванова было слышно через слово»

«Света Светина была слишком тихая»

«У Коли Николаева звук был как из бочки»

Смотрим на Discontinuity, проверяем PLC и сеть

Смотри на Loudness, проверяем параметры AGC

Смотрим на Coloration, проверяем настройки кодеков и качество обработки

NISQA-s + Палантир = 💙

- В текущей реализации нет мониторинга качества
- Обращения пользователей крайне расплывчаты:

«Ваню Иванова было слышно через слово»

«Света Светина была слишком тихая»

«У Коли Николаева звук был как из бочки»

«Я стоял у отбойного молотка и почему-то меня никто не слышал из-за шума»

Смотрим на Discontinuity, проверяем PLC и сеть

Смотри на Loudness, проверяем параметры AGC

Смотрим на Coloration, проверяем настройки кодеков и качество обработки

NISQA-s + Палантир = 💙

- В текущей реализации нет мониторинга качества
- Обращения пользователей крайне расплывчаты:

«Ваню Иванова было слышно через слово»

«Света Светина была слишком тихая»

«У Коли Николаева звук был как из бочки»

«Я стоял у отбойного молотка и почему-то меня никто не слышал из-за шума»

Смотрим на Discontinuity, проверяем PLC и сеть

Смотри на Loudness, проверяем параметры AGC

Смотрим на Coloration, проверяем настройки кодеков и качество обработки

Смотрим на Noisiness, проверяем работу шумодава

NISQA-s + Палантир = 💙

- В текущей реализации нет мониторинга качества
- Обращения пользователей крайне расплывчаты:

«Ваню Иванова было слышно через слово»

«Света Светина была слишком тихая»

«У Коли Николаева звук был как из бочки»

«Я стоял у отбойного молотка и почему-то меня никто не слышал из-за шума»

«Все было очень плохо, не могу сказать, что конкретно»

Смотрим на Discontinuity, проверяем PLC и сеть

Смотри на Loudness, проверяем параметры AGC

Смотрим на Coloration, проверяем настройки кодеков и качество обработки

Смотрим на Noisiness, проверяем работу шумодава

NISQA-s + Палантир = 💙

- В текущей реализации нет мониторинга качества
- Обращения пользователей крайне расплывчаты:

«Ваню Иванова было слышно через слово»

«Света Светина была слишком тихая»

«У Коли Николаева звук был как из бочки»

«Я стоял у отбойного молотка и почему-то меня никто не слышал из-за шума»

«Все было очень плохо, не могу сказать, что конкретно»

Смотрим на Discontinuity, проверяем PLC и сеть

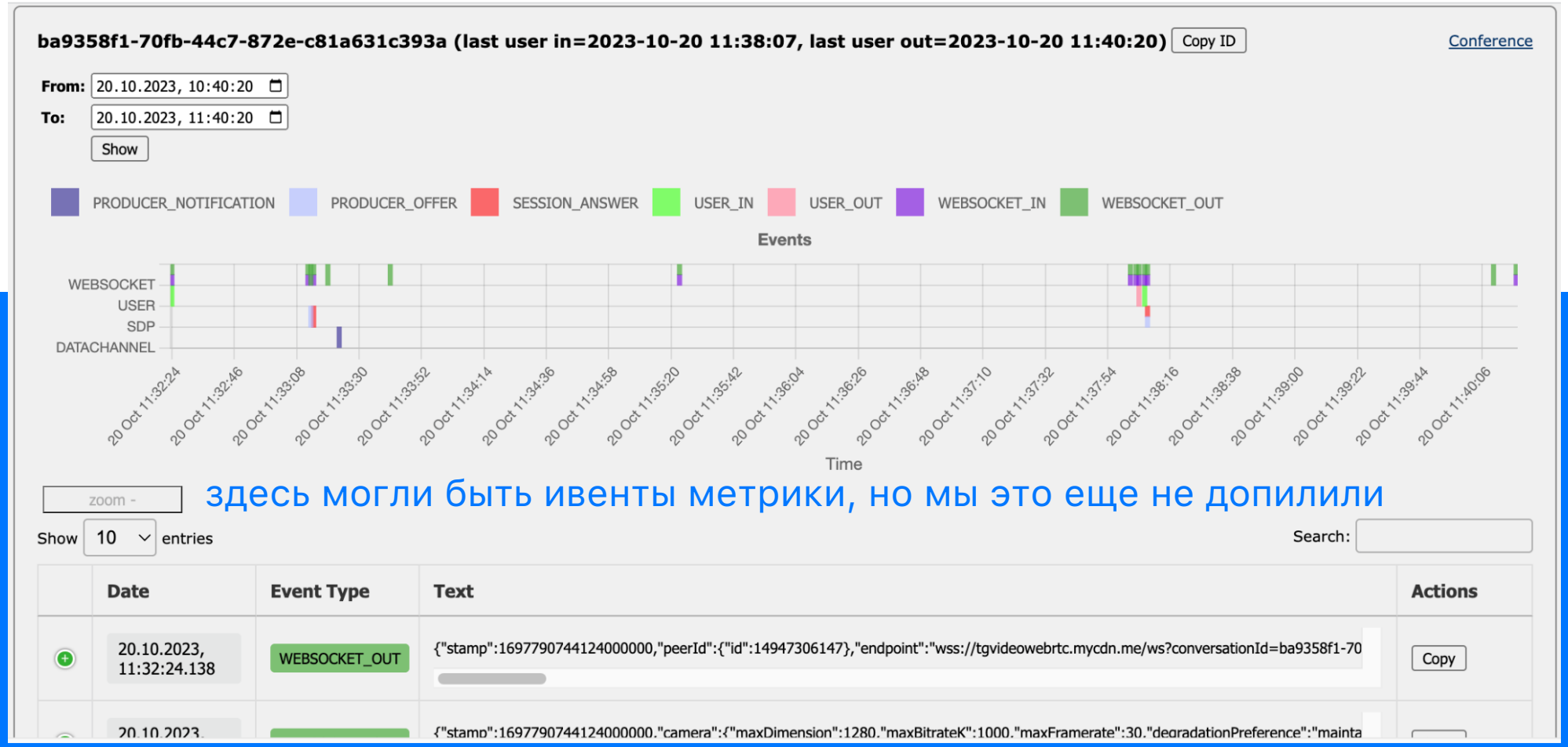
Смотри на Loudness, проверяем параметры AGC

Смотрим на Coloration, проверяем настройки кодеков и качество обработки

Смотрим на Noisiness, проверяем работу шумодава

Смотрим на MOS в конце и на метрики на каждом этапе

NISQA-s + Палантир = 💙



Что со всем этим делать?

Наши собственные планы по использованию

- Уже логгируется в дашборды на бэкенде ВК Звонков
- Интеграция в Палантир
- Проработка UI для пользователя

Ваши планы
могут быть
какими угодно!

https://github.com/deepvk/NISQA-s

- Код и подробное ридми
  для инференса и обучения
- Наш чекпойнт
- Ссылка на датасеты
- Оффлайн и онлайн инференс
- Модифицируемый конфиг
  и подробные комментарии

Спасибо за внимание!

Иван Бескровный, программист-разработчик
в Команде звуковых технологий ВКонтакте