Практическая работа №5
Trade-off и метрики релиза

Цель работы: научиться принимать решения о приоритетах (trade-off переговоры), исходя из метрик, ресурсов и неожиданных событий. Связать качество релиза с бизнес-эффективностью.

Матрица приоритетов с учётом непредвиденных факторов

Релиз 7: Внедрение AI-ассистент для автоматической категоризации и тегирования документов

Во время релиза возникли следующие "черные лебеди": Внешний AI-сервис резко поднял цены в день релиза, а также юридический отдел потребовал вести лог всех AI-решений для аудита.

Высокие затраты	Логирование AI-решений Ограничение частоты запросов	AI-категоризация Кэширование результатов AI Fallback на локальную модель
Низкие затраты		Расширенный набор категорий (30 типов) Автоматическое обучение модели Голосовая категоризация
	Низкие усилия	**Высокие усилия**

Таблица метрик релиза (TCO, CPU, ROI, Run vs Grow и др.)

Метрика	Плановое значение	Фактическое значение
TCO(Total Cost of Ownership)	4 936 000 руб.	4 936 000 руб.
CPU(Cost per Unit)	14.9 руб. (один запрос)	53,08 руб.(после увеличения на 0,5$)
ROI(Return on Investment)	11,1%	6,1%
Run vs Grow Ratio	25%/75%	40%/60%

Анализ влияния рисков и итоговый план корректировок

Негативные последствия

- Рентабельность релиза просела почти в два раза
- Рост операционных затрат увеличивает срок окупаемости
- Задержки из-за новых требований (логирование)
- Рост нагрузки на инфраструктуру
- Появление зависимости от внешнего поставщика.

Положительные сигналы

- Требования юридического отдела увеличивают доверие клиентов, следовательно повышают долгосрочный ROI
- Локальная модель и кэширование снижают будущие риски
- Проект становится более устойчивым и масштабируемым

Итоговый план корректировок

Срочные корректировки:

1. Внедрить кэширование AI-результатов

- Снижение CPU минимум на 30–60%
- Быстрая реализация (1–2 дня)

2. Ввести rate limiting

- Снижение TCO при росте цен
- Уменьшение нагрузки на API

3. Реализовать логирование AI-решений

- JSON-лог
- Хранение 90 дней.

Среднесрочные корректировки

4. Построить Fallback-модель

- Снижение зависимости от внешнего SaaS
- Использовать при превышении лимитов/авариях

5. Оптимизировать затраты:

- переход на пачки запросов
- выбор более дешёвого тарифа AI

Функции, которые следует перенести

6. Расширенный набор категорий (30 типов) - перенос на следующий квартал.

7. Автоматическое обучение модели - высокие затраты, низкий ROI сейчас

8. Голосовая категоризация - заморозить до стабилизации Run-cost