Чек-лист архитектуры Video RAG

Fora Soft Learn · AI for Video Engineering

1 · Нужен ли вообще RAG? (против модели с длинным контекстом)

- [ ] Одно умеренное видео, пара вопросов: пропустите RAG. Отправьте всё в модель с длинным контекстом. Проще.
- [ ] Архив слишком БОЛЬШОЙ: 1 час видео ≈ 947 000 токенов (~263/сек). Сточасовой архив не влезает. Нужен RAG.
- [ ] Слишком ДОРОГО перечитывать: длинный контекст перечитывает видео на каждом запросе (~$2,37/час у Gemini 2.5 Pro). RAG индексирует раз, потом читает чанки.
- [ ] Точность: модели уделяют внимание неравномерно после ~400k токенов. Короткий извлечённый контекст точнее.

2 · Выберите подход извлечения (шпаргалка)

```
Text grounding    : ASR + OCR + описания → text RAG | дешевле, зрелее; с потерями по визуалу
Визуальные эмбед. : эмбеддинг кадров/клипов модель.  | тяжелее; визуальный смысл выживает
Гибрид            : текст для массы + визуал для визуальных запросов | почти все продакшн
Правило           : в осн. СКАЗАНО/на экране → текст | в осн. ВИДНО → визуал | оба → гибрид
```

3 · Индексация — один раз на видео (несёт стоимость)

- [ ] Режьте по СЦЕНЕ/говорящему/теме, никогда по фиксированному интервалу. Пусть чанки слегка перекрываются.
- [ ] Отбор кадров до обработки: даунсемплинг 60→4 fps, затем только ключевые кадры (~90× меньше).
- [ ] Извлеките: ASR (с таймкодами) + OCR + описания сцен через VLM, или визуальные эмбеддинги — по выбранному подходу.
- [ ] Индексируйте эмбеддинг каждого чанка в vector DB (Pinecone/Weaviate/Milvus/Qdrant) С таймкодами + метаданными.

4 · Запрос — на каждый вопрос (дёшево и с доверием)

- [ ] Эмбеддинг вопроса на ту же карту смысла; берём top-k (3-10) ближайших чанков из vector DB.
- [ ] Reranking короткого списка кандидатов до лучших 2-3 перед генерацией — дёшево, список уже короткий.
- [ ] Генерируйте ТОЛЬКО по найденным чанкам и показывайте таймкоды + исходные клипы — проверка человеком в один клик.
- [ ] Форма стоимости: платите раз за индекс; каждый вопрос читает пару тысяч токенов (<1 цента). Разрыв растёт с масштабом.