

1 — Цикл ревью как флот (цикл агента, наведённый на весь архив)

Архив -> очередь работ -> N параллельных воркеров, каждый: восприятие -> анализ -> действие -> запись
Fan-out: 1000 воркеров осушают линию в 1000 раз быстрее. Часов нет, единого запроса нет.
Уверенные вердикты -> каталог; низкая уверенность / риск -> очередь ревью людьми (эскалация).

2 — Воронка стоимости (дешёвые инструменты вперёд; дорогой reader — на остатке)

Дешёвый фильтр	смена планов + транскрипт; убрать пустоту	(на всём)
Сэмплер кадров	взять несколько кадров на план	(низкая)
Reader (VLM)	описать выбранные кадры	(ВЫСОКАЯ — держать последним)
Судья по политике	описание -> структурный вердикт + оценка	(средняя)
Каталог / эскал.	уверенное хранить; остальное — людям	(низкая)

3 — Экономика «без часов» (никто не ждёт — оптимизируй стоимость, не скорость)

Batch-API: OpenAI / Anthropic / Gemini — все 50% за 24 ч. Real-time их не использует.
Математика воронки: 18 000 кадров/видео -> ~40 = ~450x меньше работы reader; batch делит остаток вдвое.
Spot / preemptible: дёшево и прерываемо — безопасно ТОЛЬКО если задача переживает прерывание (см. 4).

4 — Хребет надёжности (сбой на 80 000 продолжает с 80 001, не заново)

ЧЕКПОИНТ	записать готовый элемент до перехода к следующему.
ИДЕМПОТЕНТНОСТЬ	ключ = video-N + модель + версия политики; пропускай готовое (иначе двойной счёт).
RETRY + DLQ	повтор временных сбоев; «ядовитый» элемент -> dead-letter -> один человек, не вся задача.
ОРКЕСТРАЦИЯ	durable-движок (напр. Temporal) проигрывает event-log и продолжает с середины.

5 — Чек-лист (спроси это до того, как строить, купить или обернуть)

- ☐ Есть ли воронка дешёвый фильтр + сэмплер, чтобы дорогой reader видел ~десятки кадров, а не каждый?
- ☐ Тяжёлые вызовы модели идут через batch-API (50%), потому что никто не ждёт?
- ☐ Прогресс чекпойнтится, чтобы сбой продолжал со следующего элемента, а не начинал заново?
- ☐ У каждой единицы работы есть ключ идемпотентности (видео + модель + версия политики) против дублей?
- ☐ Сбои повторяются, а очередь dead-letter изолирует «ядовитые» элементы на разбор человеком?
- ☐ Вердикты версионизируются, чтобы переобработка после смены модели/политики была частичной, а не полной?
- ☐ К людям идут только низкая уверенность и высокий риск, а случайную долю уверенных перепроверяют?
- ☐ Хранится ли причина каждого вердикта (улики + пункт политики) ради прозрачности модерации?