

1 - Две половины на двух часах (главная идея)

Индекс (один раз)	Пакетный проход по ВСЕМУ архиву по своему графику — никто не ждёт. Флот асинхронного ревью по batch-цене (–50%). Строит каталог один раз.
Расследование (по запросу)	Цикл агента срабатывает на вопрос человека. Ищет в каталоге, читает лишь важные клипы. Оптимизирован под скорость и точность, не под цену.

2 - Что хранит индекс (три слоя, от дешёвого)

Детекции + треки	CV-модели (YOLO + трекер). Счёт и движение: сколько, где, когда, как быстро. Самый дешёвый слой.
Эмбединги	Числа, кладущие клипы и запросы на одну карту смысла; ближайшее = похожее. В векторной БД. Поиск на обычном языке.
Описания	Короткие сводки от читателя (модель видение-язык), только для помеченных сегментов. Самый богатый и дорогой; «что произошло».

3 - Воронка стоимости (один запрос: 40 камер x 1 год)

~38 000 000 000 кадров в архиве	-> прочитать всё = невозможно
дешёвый фильтр + выбор кадров	-> объём меньше в сотни раз (один раз)
структурный поиск по индексу	-> ~14 кандидатов (быстрый запрос к БД)
читатель подтверждает ТОЛЬКО их	-> 14 чтений, а не 38 млрд -> 11 готово
Правило: сперва дешёвый индекс; дорогой читатель последним, на горстке.	

4 - Хребет: надёжность, наблюдаемость, оценка

Надёжность	Чекпоинты LangGraph продолжают упавшее расследование; движок надёжного исполнения (класса Temporal) защищает линию индекса. Инструменты через MCP.
Наблюдаемость	Трейсы OpenTelemetry GenAI по каждому шагу плана / инструмента / модели (LangSmith, Langfuse, AgentOps). Не можешь проиграть — не можешь доверять.
Оценка	Золотой набор на каждое изменение. Пропуски против ложных срабатываний; потолок вызовов на запрос, чтобы один разгон не сжёг бюджет.

5 - Чек-лист красных линий EU AI Act (ограничения дизайна)

- ☐ По умолчанию ОБЪЕКТЫ + СОБЫТИЯ, не личности — «погрузчик у человека», а не «это Джейн Доу».
- ☐ Постфактумную биометрическую идентификацию — вне пути по умолчанию (высокий риск; правоохранителям нужна судебная санкция).
- ☐ Без удалённой биометрической идентификации в реальном времени в публичных местах (запрет с 2 февраля 2025).
- ☐ Раскрывайте взаимодействие с AI и AI-сгенерированный / изменённый контент (прозрачность, Статья 50).
- ☐ Контроль человека на каждый значимый вывод — агент предлагает, человек распоряжается.
- ☐ Логируйте, кто что решил; держите аудит-трейл. Это не юридический совет — зовите юриста к конкретике.