Video Archive Search & Q&A Engine - Build Blueprint

Fora Soft Learn · AI for Video Engineering · Capstone 10.5

1 - The reference architecture: two halves, two rules

Rule 1	Index once, retrieve per question - the archive never enters the model.
Rule 2	No answer without a citation to the source clip and timecode.
Ingestion	Chunk on scenes -> select keyframes -> extract (ASR+OCR+VLM) -> embed -> index (DB + timecode + perms).
Query	Understand -> hybrid retrieve (semantic + keyword) -> rerank -> answer model reads only the few chunks -> cite.

2 - Build vs buy (2026): rent the models + vector DB, build the rest

```
BUY / HOST  embedding model + answer model (VLM/LLM) + ASR/OCR/captioner - fastest-moving parts
BUY         vector database (Pinecone / Qdrant / Weaviate / Milvus) - solved infra, do not reinvent
BUILD       extraction pipeline + hybrid retrieval + citation surface + governance = your product
NOTE        Twelve Labs sunset Marengo 2.7 on 30 Mar 2026 - that is why you abstract the embedding model.
```

3 - The 2026 model field (verify before you ship)

```
EMBEDDING (expensive swap): Gemini Embedding 2 (multimodal 3072-dim) | Cohere Embed 4 | Marengo 3 | Nova
  open: SigLIP 2 | Jina v5 | Qwen3-Embedding   - changing it INVALIDATES every vector -> re-index.
ANSWER (cheap swap): Gemini 2.5 | Claude | GPT | open LLaVA / Qwen-VL - reads only a few chunks.
Keep the extracted text (ASR/OCR/captions) beside the vectors so a re-index re-embeds from text.
```

4 - The two-line cost model (keep these on separate budget lines)

```
ONE-TIME INDEX  ~$2.50 / hour of video (ASR ~$0.36 + captioning & embedding ~$1-2) + ~cents/mo storage.
PER QUESTION    ~$0.01-0.02 (embed + vector search + rerank + generate over a few retrieved chunks).
Index 1,000 hrs once ~= $2,500 fixed, then ~1-2c per question.
Long-context baseline: 1 hour ~= 946,800 tokens x $2.50/M = ~$2.37/question; a full archive does NOT fit.
```

5 - Build order: each milestone ships a working tool

```
1 Cited search box (text-grounding index + clips that deep-link to the exact second).
2 Grounded Q&A with abstention (answer only from retrieved chunks; cite every claim) - second, NOT last.
3 Hybrid retrieval + reranking (add the keyword half; fix the retrieval-miss failure mode).
4 Governance (permissions + privacy + audit).   5 Scale, re-indexing, retrieval evaluation.
```

6 - Go-live + governance checklist (engineering context, not legal advice)

- [ ] Every chunk carries source video + timecode from ingestion, so every answer can cite (Rule 2).
- [ ] Embedding model behind one interface + extracted text retained, so a model retirement is a scheduled re-index.
- [ ] Hybrid retrieval (semantic + keyword) + reranking - exact names/codes and meaning both caught.
- [ ] Permissions filtered at RETRIEVAL time by the asker's current rights - never just at index time.
- [ ] Person-identity search treated as higher-risk (GDPR + EU AI Act biometric) - often a decision NOT to build.
- [ ] Audit log of every question + clip returned; retrieval evaluated against known-answer questions, not just prose.