1. Text Embedding Models (Open Source)

Model	Type	Notes
Instructor XL	Embedding	Text + task-aware context
E5 (text2vec)	Embedding	Good multilingual, very fast
GTE-base / large	Embedding	Compact, multilingual, good recall
BGE / BAAI	Embedding	Optimized for semantic search

2. Image Models (Vision Encoders)

Model	Base	Notes
CLIP (OpenAI)	Vision + text	Gold standard for image-text embeddings
OpenCLIP	CLIP variant	Trained on LAION datasets, open weights

BLIP / BLIP-2	Vision + text	Captioning + QA, great for scanned docs
LLaVA	Vicuna + CLIP	Visual chat is good for document screenshots

Use these to extract **image features** or **caption images** to convert into searchable text.

3. Audio & Speech Models

Model	Function	Notes
Whisper (OpenAI)	Speech-to-text	Best open-source STT, multilingual
WhisperX	Whisper + timestamping	Adds speaker diarization + word time
Wav2Vec 2.0	Audio embedding	Use for audio similarity

Transcribe and chunk transcripts with **timestamps** → store them in a vector DB.

4. Video Handling (Hybrid Approach)

- **Frame Extraction**: Use `ffmpeg` to extract keyframes (every X seconds or scene changes)

- **Audio Transcription**: Run Whisper on video audio
- **Frame Embedding**: Use CLIP or BLIP to get frame-level embeddings
- **Index in DB**: Timestamped image + text pairs

5. Vector Databases (Search Infrastructure)

Name	**Features**
FAISS	Fastest, simplest ANN search
Weaviate	Schema-based, multimodal search
Qdrant	Filters, payload, Docker-friendly
ChromaDB	Easy to use, simple APIs

6. Optional: Multimodal LLMs (Inference)

Model	**Base**	**Use case**
MiniGPT-4	Vicuna + BLIP	Visual+text Q&A, captioning

LLaVA One Vision 7B	Vicuna + CLIP	Visual instruction following
BakLLaVA	LLaMA 3	Open-weight, vision+text
Fuyu	Deformable DETR	Newer vision model from Adept
CogVLM	Stronger multimodal	Better visual grounding
Qwen 2.5 72B		

These can help **summarize**, **explain**, or **generate search previews**, but are optional for pure search.

Role of the LLM in Semantic Search App

1. Query Understanding

- **Paraphrasing or expanding user queries** to improve semantic matching.
- Example: User types “how to build a hollow part?” → LLM rephrases to match:
  “What sequence design and primitives are used to create a hollow axisymmetric component in NAGFORM?”

2. Multimodal Embedding Generation

- If using **LLaMA 3.2 Vision or LLaVA**, the LLM includes **built-in encoders** for:
    - Text embeddings (from queries and documents)
    - Image understanding (e.g., diagrams in PDFs or screenshots)
    - Basic video/audio captions

  Alternative: Can use **separate encoders** (e.g., CLIP for vision, Whisper for audio)

3. Retrieval-Augmented Feedback

- After the vector search retrieves top matches, the LLM can:
    - **Summarize** retrieved snippets
    - **Rank or cluster** results
    - Provide user-friendly **contextual previews** without inventing facts

4. Multimodal Reasoning

- For use cases like:
    - Understanding a **diagram with captions**
    - Explaining a video segment with **text + visual** info
    - Relating **text and image in a single paragraph**

  If needed, you can use LLaVA, MiniGPT-4, or GPT-4o for this stage.

5. UI Co-Pilot (Optional Chatbot Layer)

- The LLM can serve as a **semantic interface layer**:
    - Interpret complex user queries
    - Keep the multi-turn conversation context

- Provide "Ask me anything about NAGFORM manuals" UX