Name : Sai Sruthisri
Date : 3/12/25

1) Core LLM (API models)

Requirements:

- Scalable
- Actively maintained
- Very low probability of abrupt deprecation
- Strong instruction-following
- Multilingual capability
- Long-term (2+ year) reliability

1) LLAMA (docs)
   Models we can use : Llama 3.3 (70B , 8B)
   Depcapriation : No date mentioned (So not anytime soon)
   Tokens :128K
   Data type : Only text
   Multilingual Support: Yes (100+)

2) Qwen3-Coder 30B A3B Instruct (docs)
   Depcapriation : No date mentioned (So not anytime soon)
   Release Date : August 2025
   Tokens :262,144 tokens
   Data type : Sparse Mixture-of-Experts code-generation / instruct model from Qwen3 series — optimized for large context, repository-scale code understanding and complex code tasks.Strong code-generation, tool-calling, agent-style

   Multilingual Support: Yes a little lesser than other models

3) Mistral-7B-Instruct-v0.3 (docs)
   Depcapriation : No date mentioned (So not anytime soon)
   Release Date : July 2024
   Tokens :32.8K tokens
   Data type : Highly efficient 7B model known for strong performance and instruction following
   Multilingual Support: No

4) gpt-oss-120b(docs)
   Depcapriation : No date mentioned and it the latest model as of now (So not anytime soon)

Release Date : 2024
Tokens : 31.1k tokens
Data type :OpenAI's largest open-weight reasoning model designed for powerful reasoning, agentic tasks, and versatile developer use cases.
Multilingual Support: Yes but quality is maybe not confimly very good

5) DeepSeek R1 Distill LLaMA 70B(docs)
Depcapriation : No date mentioned (So not anytime soon)
Release Date : January 2025
Tokens : 64k tokens
Data type :DeepSeek R1 Distill LLaMA 70B is a distilled, efficient version of DeepSeek's R1 reasoning model. It provides strong chain-of-thought reasoning, advanced problem solving, and high-quality text generation with significantly reduced GPU memory requirements
Multilingual Support: Maybe no cause it is for chain-of-thoughts, mathematics , logics

6) DeepSeek R1 (docs)
Depcapriation : No date mentioned , released this year only (So not anytime soon)
Release Date : January 2025
Tokens : 128K tokens (32K–64K usable actually through hugging face)
Data type :DeepSeek R1 excels at **advanced reasoning**, especially in math, logic, and multi-step problem solving.It also performs well in coding, general chat, and tool-enabled tasks, but reasoning is its main strength.
Multilingual Support: **Yes**
Strength Levels:
Strong: English, Chinese
Good: Hindi, Japanese, Korean, French, Spanish, German
Moderate: Other European & Indian languages
Weak: Very low-resource languages
It works best in English + Chinese.

2) Vector Embeddings

Model / Framework	What it offers / Why good
LaBSE	Provides language-agnostic sentence embeddings across 100+ languages; open-source under Apache 2.0. Good for cross-lingual retrieval / embedding.

Sentence-Transformers (plus multilingual models)	A widely used framework; many pre-trained models (multilingual) available under permissive license, giving you dense sentence embeddings out-of-box.
GTE Multilingual Base	A newer model supporting 50+ languages, explicitly designed for multilingual sentence embeddings / dense retrieval, released under Apache-2.
(Alternative) **LASER**	Older but proven multilingual sentence-embedding toolkit; supports dozens of languages with a unified embedding space.

3) Vector Databases

1) Qdrant : embedding PDFs, storing chunks, needing metadata + semantic search, building a retrieval/query + re-embedding pipeline. Qdrant is often the "go-to" for mid-sized vector stores with metadata + good performance.

2) ChromaDB : Early-stage/new projects, prototyping RAG pipelines, moderate-sized document collections (e.g. a few thousand to maybe 100k chunks), or when you just want something simple and easy to integrate with Python.

3) Milvus : If we need scaling up to very large document stores (millions of documents / embeddings), need high throughput and low latency for many concurrent users/queries — e.g. enterprise knowledge bases, large-scale retrieval + RAG systems, multimodal data stores, etc.

4) Pinecone

- **Type:** Fully managed vector database
- **Free Tier:** Yes, ~1M vectors free + 512 dimension per vector (as of 2025)
- **Scalable:** Serverless & fully managed
- **Multilingual/embedding model support:** Works with *any embedding model* (you push vectors yourself)
- **License / API:** API access, cloud-hosted, no local hosting needed
- **Limitations:** Free tier limited in vectors and queries; larger scale requires paid plan