**Vector Database Challenge Set**

Each student must:

* **Generate their own dataset** (≥2000 records, stored in CSV/TXT/DOCX/PDF).
* **Process through a pipeline** (load → clean → chunk → embed → store in FAISS/Qdrant).
* **Build a retrieval system** for the challenge scenario.
* **Demonstrate with your demo video in group .**

**Challenge Scenarios**

1. **Multilingual Knowledge Base**
   * Store English + Hindi/Hinglish content.
   * Query in Hindi, retrieve English documents.
   * Bonus: test dual-index vs multilingual embeddings.
2. **Legal Document Similarity Checker**
   * Upload mock contracts/policies.
   * Query for clauses (“termination conditions”) and find semantically similar sections.
3. **Customer Support Agent**
   * Store FAQs and past support tickets.
   * Build a retrieval-based chatbot that suggests answers.
4. **E-commerce Product Search**
   * Store product descriptions as embeddings.
   * Support “search by natural language” and “search by image” (if using CLIP).
5. **Academic Research Helper**
   * Ingest academic papers or abstracts.
   * Retrieve papers semantically for queries like “transformers in computer vision.”
6. **Code Snippet Retrieval**
   * Store Python/Java functions.
   * Retrieve by query (“binary search in Python”).
7. **Movie/Book Recommendation Engine**
   * Store plot summaries/reviews.
   * Recommend nearest neighbors based on user preferences.
8. **Resume–Job Matching**
   * Store resumes and job descriptions.
   * Retrieve top matches per job posting.
9. **Plagiarism Detector**
   * Store student essays.
   * Check new submissions for high-similarity overlaps.
10. **Personalized News Feed**
    * Store embeddings of news articles.
    * Compare to user profile (based on liked articles).
11. **Healthcare Diagnostic Assistant**
    * Store symptoms → disease mappings.
    * Query with user-described symptoms in plain text.
12. **Sentiment Cluster Explorer**
    * Store tweets/reviews.
    * Cluster embeddings, visualize with t-SNE/PCA, analyze sentiment groups.
13. **Music/Lyrics Search Engine**
    * Store lyrics dataset.
    * Retrieve songs by mood queries (“romantic sad song”).
14. **Image Similarity Search**
    * Store image embeddings.
    * Retrieve visually similar images.
15. **Educational Quiz Builder**
    * Store textbook passages.
    * Retrieve relevant chunks and auto-generate MCQs.