Vector DB evaluation part

Evaluation of the Vector DB & Embedding Module in Your Research Project

Since you're fine-tuning an **embedding model** on **Sri Lankan legal texts**, handling **multilingual data**, and involving **domain experts (lawyers) for evaluation**, your approach should cover both **quantitative** (accuracy, relevance) and **qualitative** (expert validation) aspects. Below is a structured evaluation plan for your module.

---

1. Evaluation Framework

Your evaluation should focus on:

1. **Embedding Model Performance** – How well the model generates meaningful vector representations.
2. **Retrieval Accuracy** – How relevant are the retrieved legal cases to the input query?
3. **Multilingual Handling** – Does the model maintain consistency across languages?
4. **Legal Expert Validation** – Do lawyers find the retrieved cases useful and accurate?

---

2. Input, Output, and Process for Evaluation

✅ Input:

- A set of **legal queries** (questions users might ask).
- A **gold standard dataset** of **Sri Lankan legal case-law and acts** (with expected correct answers).
- Multilingual queries to check language consistency.

✅ Output:

- **Ranking of retrieved cases** based on similarity scores.
- **Quantitative scores** (e.g., precision, recall, MRR).

- **Qualitative feedback** from legal experts on relevance.

✅ Process:

Step 1: Fine-Tuning the Embedding Model

- Train the model on **Sri Lankan case-law** using **contrastive learning** (if AngelBERT) or traditional fine-tuning methods.
- Ensure support for **Sinhala, Tamil, and English legal texts**.

Step 2: Legal Query Testing

- Use a set of **predefined legal queries** to test retrieval accuracy.
- Compare different embedding models (e.g., **AngelBERT, LegalBERT, MiniLM**).

Step 3: Retrieval & Scoring

- Store legal case embeddings in **Pinecone (vector database)**.
- For each **test query**, retrieve the **top 5-10 most relevant cases**.
- Measure **cosine similarity** between query and retrieved cases.

Step 4: Automatic Evaluation Metrics

- Use standard **IR (Information Retrieval) metrics**:
    - **Precision@k** – % of relevant cases in top-k retrieved results.
    - **Recall@k** – How many relevant cases were retrieved.
    - **Mean Reciprocal Rank (MRR)** – How early in the list is the correct case.
    - **Normalized Discounted Cumulative Gain (NDCG)** – Measures ranking quality.

Step 5: Legal Expert Evaluation

- Ask **lawyers to rate** retrieved cases (on a **scale of 1-5** based on relevance).
- Compare **model rankings vs. expert rankings** using:

- **Spearman's Rank Correlation** – Measures agreement with experts.
- **Fleiss' Kappa** – Checks consistency among multiple experts.

Step 6: Model Comparison & Selection

- Compare **different embedding models** (LegalBERT, AngelBERT, etc.).
- Select the best model based on **highest retrieval accuracy & expert approval**.

---

3. How to Document This in Your Research Report

📝 Section 1: Experimental Setup

- Explain **why** retrieval accuracy matters in legal search.
- Describe **the dataset** (Sri Lankan case law + acts).
- Mention the **embedding models** you are testing.
- Describe the **vector database setup (Pinecone)**.

📝 Section 2: Evaluation Metrics & Methodology

- Define **Precision@k, Recall@k, MRR, NDCG**.
- Explain how **legal experts will validate results**.
- Provide details on **multilingual testing**.

📝 Section 3: Results & Analysis

- Show **quantitative results (tables, graphs of scores)**.
- Compare **different embedding models**.
- Provide **expert feedback summaries**.
- Discuss **which model performed best and why**.

---

4. Expected Challenges & Solutions

Challenge	Solution
Legal texts are long	Use **Longformer/BigBird** embeddings if needed.
Multilingual embedding issues	Fine-tune on **Sinhala, Tamil, and English legal texts**.
Expert evaluation is subjective	Use **multiple experts & statistical validation**.
Computational cost of fine-tuning	Use **efficient models like MiniLM or AngelBERT** if needed.

Final Thoughts

This evaluation plan will ensure that your **Vector DB & Embedding Module** is tested rigorously. By combining **quantitative IR metrics** with **expert validation**, you can scientifically prove which embedding model works best.

Let me know if you need help designing **test queries or evaluation scripts**! 🚀