***Assignment 2***

1. ***Dataset Selection***

<https://huggingface.co/datasets/isaacus/open-australian-legal-qa>

or maybe this one:

<https://huggingface.co/datasets/ibunescu/qa_legal_dataset_train>

1. ***Chunking Strategy***

* Legal texts require special chunking considerations:
  + Top-level chunking by case document (preserving full case context)
  + Mid-level chunking by legal reasoning sections (facts, arguments, holding)
  + Paragraph-level chunking for specific legal points
* Citation Preservation:
  + Maintain Australian citation format (e.g., "Nasr v NRMA Insurance [2006] NSWSC 1918")
  + Cross-reference related cases within the same jurisdiction
* Metadata Enrichment:
  + Include jurisdiction data (New South Wales, etc.)
  + Add document type classification (Decision, Order, etc.)
  + Preserve URLs to original sources
* Chunk Size Optimization:
  + Case holdings: 300-500 tokens
  + Factual backgrounds: 400-600 tokens
  + Legal reasoning sections: 500-700 tokens

1. ***Vector Database - Chroma***
2. ***Model Choice*** 
   1. ***Embedding – bge-large-en-v1.5***
   2. ***LLM – Mistral-7B-Instruct-v0.2 / Phi4-mini***
3. ***RAG Strategies***
4. Basic RAG
   1. will retrieve the most similar legal documents to a user query and use them as context for our LLM
5. Query Expansion with Legal Terminology
   1. Identify specialized legal terms in user queries
   2. expand queries with legal synonyms and related concepts
   3. add jurisdiction-specific terminology when applicable
6. Multi-Query RAG for Legal Perspectives
   1. generate 2-3 different versions of the same query to capture different aspects
   2. retrieve different document sets for each query version
   3. combine and deduplicate results before generation
7. Metadata-Enhanced Retrieval
   1. Jurisdiction filtering (Federal, NSW, Victoria, Queensland, etc.)
   2. Court hierarchy weighting (prioritize higher Australian courts)
   3. Recency analysis (Australian law evolves; newer cases may supersede older ones)
   4. Citation network mapping within Australian legal system

***Timeline:***

***Week 7-8: Foundation Building***

* Set up vector database (Chroma) and embedding pipeline
* Implement initial data ingestion framework for legal datasets
  + Create chunking logic for Australian case decisions (facts, reasoning, holding)
  + Develop metadata schema capturing Australian jurisdictions (NSW, Federal, etc.)
* Develop basic RAG pipeline with simple query-document retrieval
* Set up evaluation framework to measure performance
* Develop first prototype with basic user interface

***Week 9-10: RAG, RLHF & Hallucination Mitigation***

* Implement advanced RAG strategies:
  + Query expansion with legal terminology
  + Multi-query approach for legal perspectives
  + Metadata-enhanced retrieval
* Begin task-specific fine-tuning for legal Q&A
  + Collect small dataset of legal QA pairs
  + Implement fine-tuning pipeline for the model
* Create citation handling system
* Implement lightweight RLHF approach:
  + Create preference dataset for legal responses
  + Train reward model for legal accuracy
  + Implement feedback mechanism
* Develop hallucination detection and mitigation:
  + Citation verification against case database
  + Implement confidence scoring for jurisdiction-specific claims
  + Cross-check legal reasoning against Australian precedents
* Final testing