1. Survey một số model document retrieval

Mô hình	Kiến trúc	Phương pháp	Độ chính xác	Y/c tài nguyên	Y/c dữ liệu huấn luyện	Ưu điểm	Nhược điểm	Ứng dụng	Độ phức tạp
BERT	**Mạng nơ-ron sâu**: Kiến trúc Transformer, với nhiều lớp mã hóa và cơ chế chú ý (attention mechanism) cho phép xử lý ngữ nghĩa trong ngữ cảnh.	**Phương pháp**: Huấn luyện mô hình trên một tập dữ liệu lớn bằng cách sử dụng kỹ thuật masked language modeling (MLM) và next sentence prediction (NSP). Thực hiện fine-tuning trên dữ liệu pháp lý tiếng Việt để cải thiện khả năng truy vấn.	Rất cao	Cao	Cần nhiều dữ liệu	Hiểu ngữ nghĩa tốt, khả năng trả lời câu hỏi chính xác, có thể điều chỉnh cho các nhiệm vụ cụ thể.	Cần nhiều tài nguyên tính toán và thời gian huấn luyện.	Tìm kiếm tài liệu luật, hỗ trợ pháp lý, chatbot, phân tích văn bản luật.	Cao
Legal-BERT	Phiên bản của BERT, được điều chỉnh cho ngữ cảnh pháp lý, sử dụng các lớp Transformer với trọng số được tối ưu hóa cho văn bản pháp lý.	Huấn luyện mô hình trên một tập dữ liệu lớn gồm các văn bản pháp lý để cải thiện độ chính xác trong các tác vụ liên quan đến luật pháp. Sử dụng các phương pháp fine-tuning để tối ưu hóa cho	Rất cao	Cao	Cần tập dữ liệu pháp lý lớn	Được tối ưu hóa cho văn bản pháp lý, cải thiện độ chính xác	Cần tập huấn luyện chất lượng cao và dữ liệu phong phú.	Tìm kiếm tài liệu pháp lý, phân tích quy định, hỗ trợ pháp lý.	Cao

		các nhiệm vụ cụ thể.				cho các truy vấn và phân tích.			
DeBERTa	Cải tiến từ BERT với kiến trúc có phần chú ý (attention) mới và kỹ thuật masked language model.	Huấn luyện trên dữ liệu lớn, sử dụng kỹ thuật chú ý để cải thiện khả năng hiểu ngữ nghĩa và ngữ cảnh.	Rất cao	Cao	Cần nhiều dữ liệu để huấn luyện	Tốt cho các tác vụ ngôn ngữ phức tạp, cải thiện độ chính xác trong các bài toán.	Cần tài nguyên lớn cho huấn luyện.	Tìm kiếm và phân loại tài liệu pháp lý, phân tích ngữ nghĩa.	Cao
ColBERT v2	Mạng nơ-ron sâu, cải tiến từ ColBERT	Mã hóa truy vấn và tài liệu, cho phép tìm kiếm nhanh chóng trên tập dữ liệu lớn.	Rất cao	Cao	Cần tập dữ liệu lớn	Tìm kiếm nhanh và hiệu quả, tối ưu hóa độ chính xác.	Yêu cầu kỹ thuật cao trong triển khai.	Tìm kiếm tài liệu pháp lý, phân tích thông tin.	Cao
Splade	Mạng nơ-ron sâu, mã hóa vector sparse	Sinh ra vector sparse từ văn bản, giúp so sánh và tìm kiếm nhanh hơn.	Cao	Trung bình	Cần dữ liệu đa dạng	Tương thích tốt với hệ	Cần điều chỉnh cho ngữ cảnh	Tìm kiếm tài liệu pháp lý,	

						thống tìm kiếm hiện tại.	pháp lý.	phân tích nội dung .	
Contriever	Mạng nơ-ron sâu, mã hóa truy vấn và tài liệu	Học cách so sánh và phân tích nội dung giữa truy vấn và tài liệu.	Cao	Cao	Cần tập dữ liệu lớn	Cải thiện khả năng tìm kiếm thông qua hiểu ngữ nghĩa .	Cần tài nguyên tính toán lớn.	Tìm kiếm tài liệu pháp lý, phân tích thông tin.	Cao

2. Đánh giá kết quả đầu ra bài toán LEGAL DOCUMENT RETRIEVAL

- Kết quả sẽ được đánh giá dựa trên chỉ số MRR@10 (Mean Reciprocal Rank). Chỉ số MRR@10 tập trung vào việc đánh giá 10 kết quả đầu tiên từ hệ thống truy vấn. MRR càng cao thì mô hình truy vấn càng chính xác.
- Trong đó, nhãn được tạo nên với các tiêu chí như sau:
    - Nhãn được tạo trên cơ sở tài liệu pháp luật tiếng Việt
    - Nhãn bao gồm cả ký tự hoa và ký tự thường
    - Nhãn có thể chứa các ký tự đặc biệt như dấu câu và các ký tự pháp lý phổ biến
    - Nhãn không phân biệt dấu cách giữa các từ
- Công thức:

$$\text{MRR} = \frac{1}{U}\sum_{u=1}^{U}\frac{1}{rank_i}$$

Trong đó:

- *rank*: Vị trí của kết quả đúng đầu tiên trong danh sách kết quả.
- U: Tổng số truy vấn.