



CS336

# LEGAL DOCUMENT RETRIEVAL

## BKAI

Đinh Nhật Trường – 22521575

Phạm Huỳnh Nhật Tân – 22521309

Phạm Nguyễn Anh Tuấn – 22521610

# Outline

- 01 **DATASET OVERVIEW**
- 02 **PROPOSE METHOD**
- 03 **EXPERIMENTAL & DEMO**

# Outline

01

**DATASET OVERVIEW**

02

**PROPOSE METHOD**

03

**EXPERIMENTAL & DEMO**

# Dataset overview

01

## Dataset

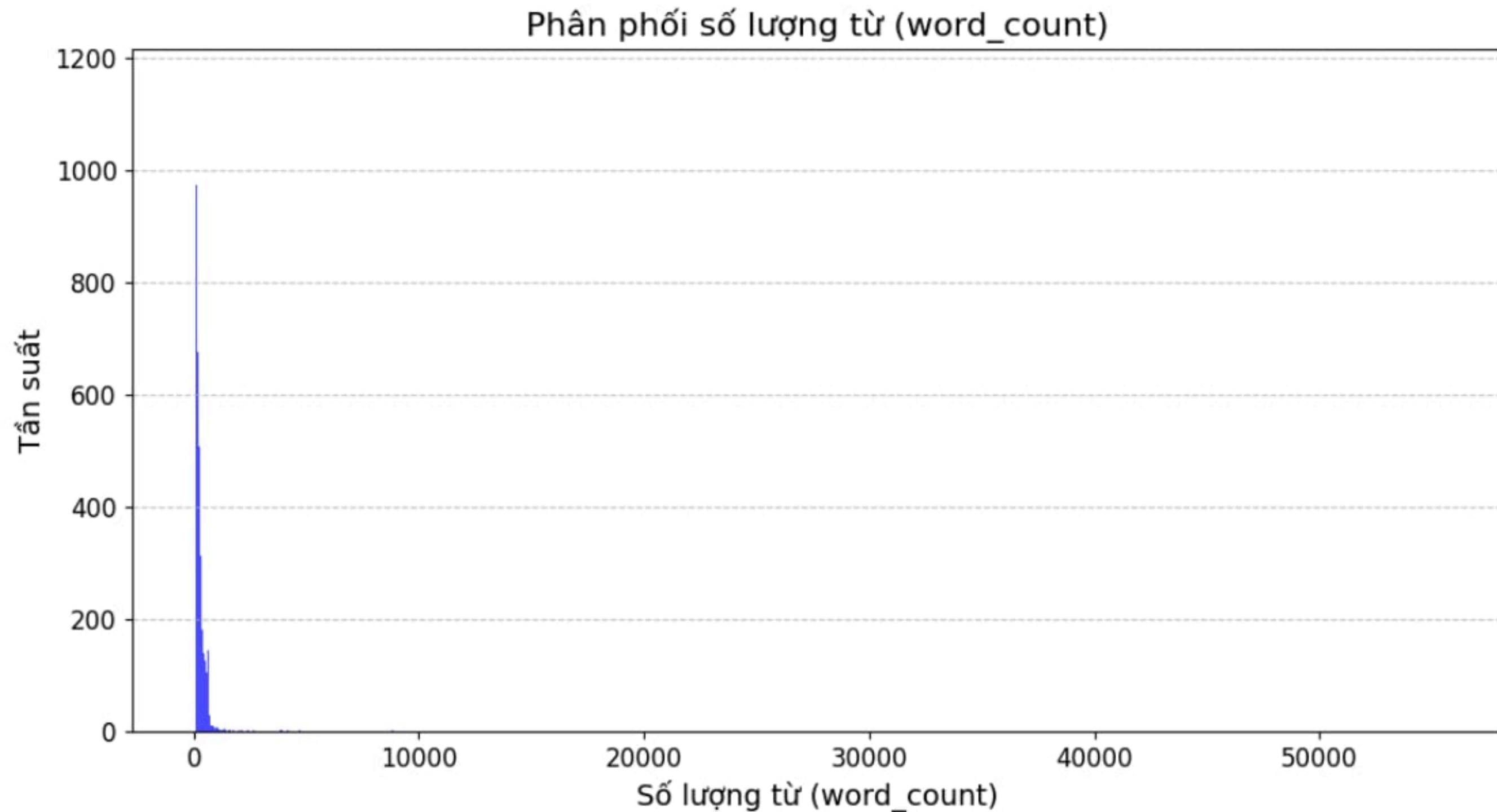
Bộ dataset pháp luật được cung cấp bởi BKAÍ gồm tập train và corpus

	text	cid
0	Thông tư này hướng dẫn tuần tra, canh gác bảo ...	0
1	1. Hàng năm trước mùa mưa, lũ, Ủy ban nhân dân...	1
2	Tiêu chuẩn của các thành viên thuộc lực lượng ...	2
3	Nhiệm vụ của lực lượng tuần tra, canh gác đề\n...	3
4	Phù hiệu của lực lượng tuần tra, canh gác đề\n...	4

# Dataset overview

02

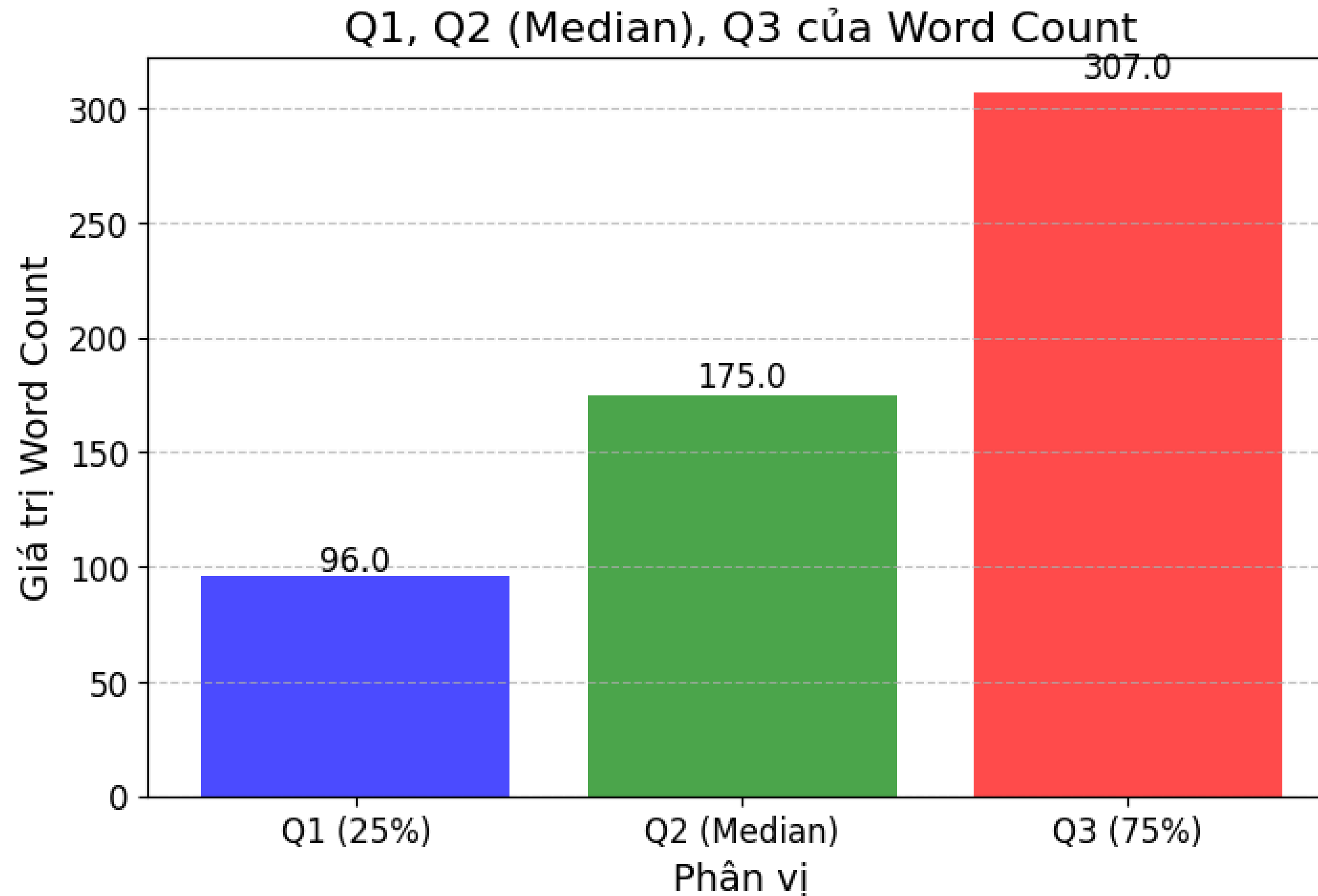
## Phân phối số lượng từ



# Dataset overview

04

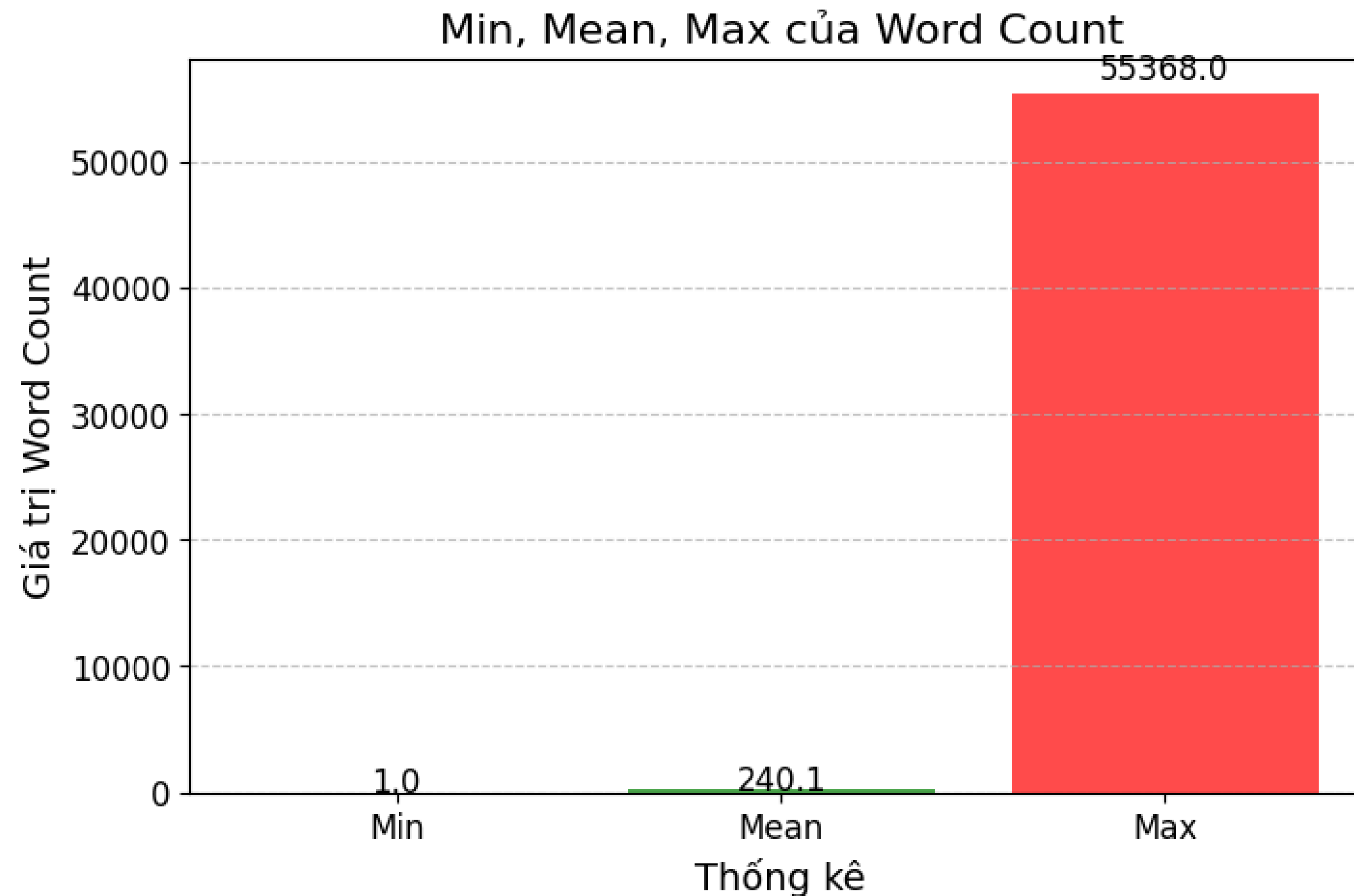
## Phân Vị



# Dataset overview

03

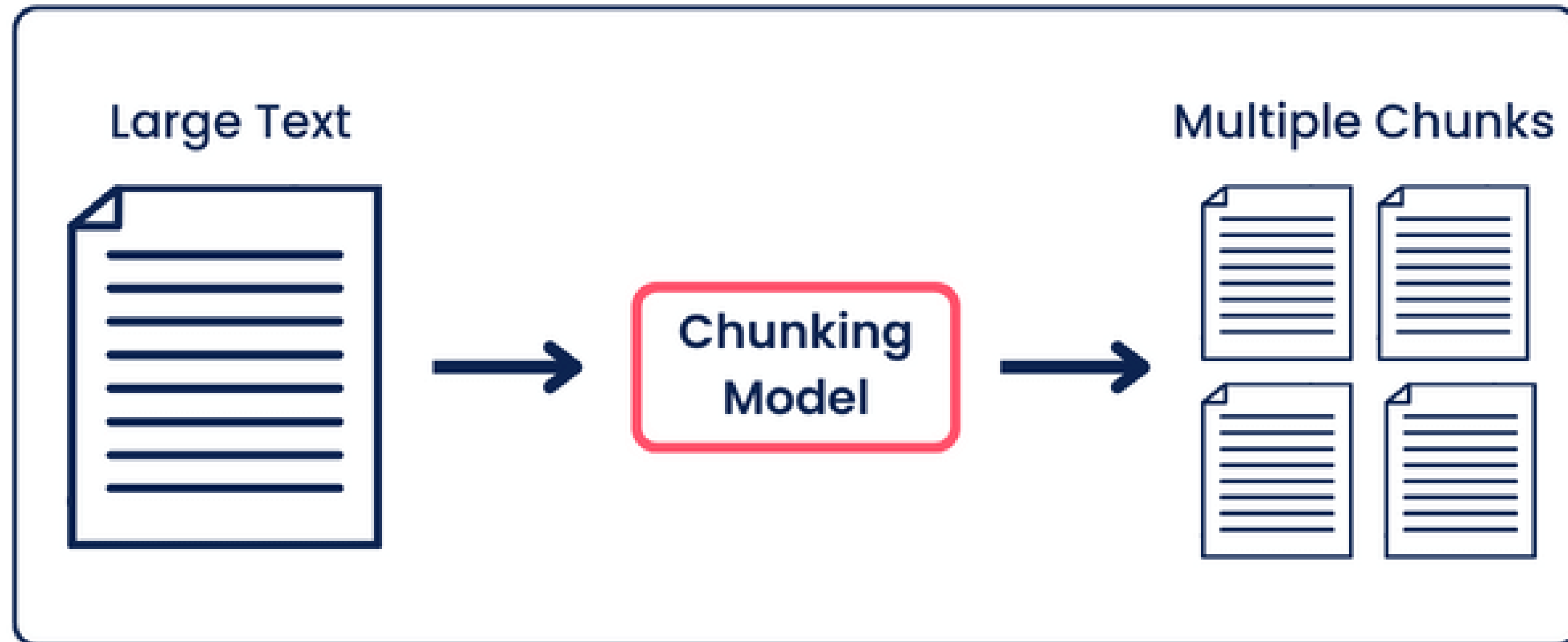
## Min, Mean và Max



# Dataset overview

## 04 Overlap Chunking

Là một trong những ứng dụng nổi bật, giúp chia nhỏ các văn bản dài dựa trên ranh giới ngữ nghĩa và token, giữ lại ý nghĩa của từng đoạn. `chunk_size = 1024`, `overlap_window = 128`





# Outline

- 01 **DATASET OVERVIEW**
- 02 **PROPOSE METHOD**
- 03 **EXPERIMENTAL & DEMO**

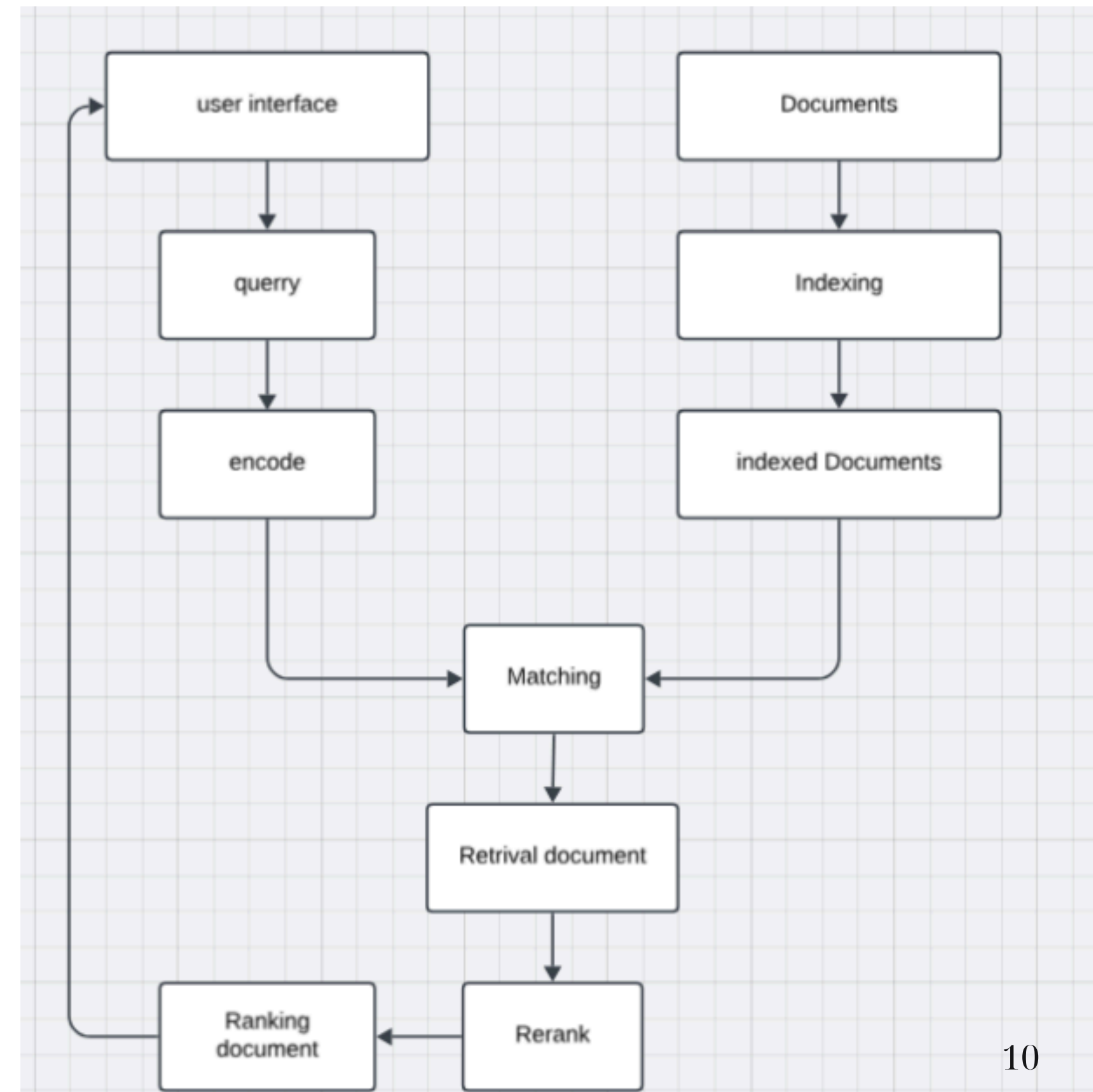
# Propose method

## 01 Overall

Kiến trúc bao gồm các thành phần chính:

- Retrieval document
- Ranking document
- Giao diện người dùng

Nó tạo điều kiện truy vấn văn bản hiệu quả thông qua mô hình và xử lý dữ liệu nhanh



# Propose method

## 2 Encoding with Jina-AI

- jina-embeddings-v3 là mô hình nhúng văn bản đa ngôn ngữ, đa nhiệm, dựa trên Jina-XLM-RoBERTa. Mô hình hỗ trợ chuỗi đầu vào dài đến 8194 token với Rotary Position Embeddings và có 5 bộ điều hợp LoRA để tạo nhúng đặc trưng hiệu quả.
- Encoding biểu diễn các chunk thành vector embedding để dễ dàng sử dụng trong tìm kiếm và so khớp ngữ nghĩa.



# Propose method

2

## Encoding with Jina-AI

Rank ▲	Model ▲	Model Size (Million Parameters) ▲	Memory Usage (GB, fp32) ▲	Embedding Dimensions ▲	Max Tokens ▲	Average ▲	ArguAna ▲	ClimateFEVER ▲
63	<a href="#">jina-embeddings-v3</a>	572	2.13	1024	8194	53.88	54.33	42.36
65	<a href="#">MUG-B-1.6</a>	335	1.25	1024	512	53.46	66.18	33.13
66	<a href="#">GIST-large-Embedding-v0</a>	335	1.25	1024	512	53.44	63.38	33.99
67	<a href="#">blade-embed</a>	335	1.25	1024	512	53.3	65.99	30.37
68	<a href="#">cde-small-v1</a>	143	0.53	768	512	53.27	72	25.71
69	<a href="#">bge-base-en-v1.5</a>	109	0.41	768	512	53.25	63.61	31.17
70	<a href="#">nomic-embed-text-v1.5</a>	137	0.51	768	8192	53.01	48.01	41.28
71	<a href="#">nomic-embed-text-v1</a>	137	0.51	768	8192	52.81	49.26	40.5
72	<a href="#">multilingual-e5-large-instruc</a>	560	2.09	1024	514	52.64	58.48	29.86

Bảng xếp hạng của Jina



# Propose method

## 3 Reranking with BGE

- Reranking là quá trình sắp xếp lại danh sách kết quả ban đầu để cải thiện độ chính xác và mức độ liên quan.
- Mô hình BAAI/bge-reranker-v2-m3 reranker nhẹ, hỗ trợ đa ngôn ngữ mạnh mẽ, dễ triển khai và có tốc độ suy luận nhanh.



# Propose method

3

## Reranking with BGE

embedding model	bge-en-v1.5 large		bge-m3		openai-small		openai-large		mxbai-embed-large-v1	
reranker	mrr	hit rate	mrr	hit rate	mrr	hit rate	mrr	hit rate	mrr	hit rate
without reranker	65.07	85.1	69.67	88.94	65.69	89.42	67.37	90.38	66.66	88.46
bge-reranker-base	75.77	90.87	77.48	94.23	75.75	93.27	76.3	94.23	76.63	91.83
bge-reranker-large	75.86	90.87	78.66	94.23	77.09	94.23	77.08	95.67	77.24	92.31
mxbai-rerank-large-v1	72.77	88.46	75.99	93.27	74.62	91.35	74.32	92.31	73.89	89.9
jina-reranker-v1-base-en	75.81	89.9	79.44	93.75	77.64	91.83	77.85	92.79	76.96	91.83
cohere rerank	75.17	90.38	76.23	91.35	76.98	92.79	76.68	93.27	76.43	92.31
ms-marco-MiniLM-L-6-v2	67.92	86.54	69.83	90.38	69.2	88.46	67.99	90.38	68.57	87.98
bge-reranker-v2-m3	78.26	90.87	80.76	94.71	79.38	93.27	79.7	94.71	79.1	92.31
bge-reranker-v2-gemma	75.19	89.9	78.14	93.75	76.74	92.31	76.28	92.31	77.25	91.83
bge-reranker-v2-minicpm-20	81.31	91.83	83.77	95.67	81.92	94.71	83.43	95.19	82.11	92.79
<b>bge-reranker-v2-minicpm-28</b>	<b>81.93</b>	<b>91.83</b>	<b>84.74</b>	<b>95.67</b>	<b>84.01</b>	<b>94.71</b>	<b>83.93</b>	95.19	<b>82.99</b>	<b>93.27</b>
bge-reranker-v2-minicpm-40	80.89	91.83	83.29	95.67	82.89	94.71	82.33	95.19	81.45	92.79

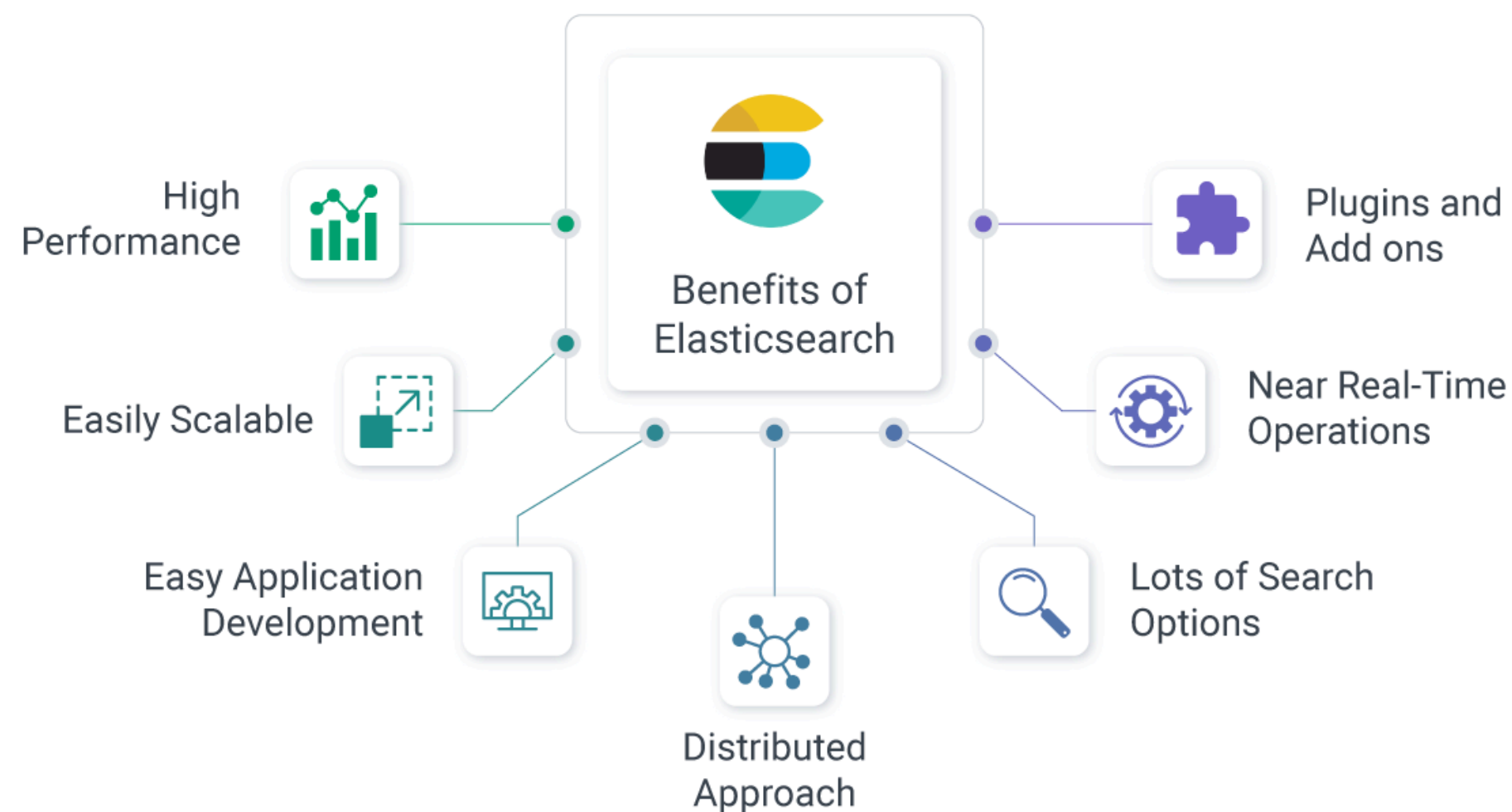
Bảng xếp hạng của BAAI/bge-reranker-v2-m3

# Propose method

4

## Vector database

**Elasticsearch** là một công cụ tìm kiếm và phân tích dữ liệu mã nguồn mở, được xây dựng trên **Apache Lucene**. Nó được thiết kế để tìm kiếm và phân tích dữ liệu văn bản, số, và dữ liệu có cấu trúc hoặc phi cấu trúc một cách nhanh chóng và hiệu quả.





# Outline

- 01 **DATASET OVERVIEW**
- 02 **PROPOSE METHOD**
- 03 **EXPERIMENTAL & DEMO**



# Experimental & Results

## 01 Experimental data

- Bộ data test được trích từ 9456 sample cuối của tập train và tập train chưa trích có tổng cộng 119456 sample
- Độ đo tính theo MRR@10 được tham khảo bởi cuộc thi BKAI \_2024

	question	context	cid	qid
0	Người học ngành quản lý khai thác công trình t...	[Khả năng học tập, nâng cao trình độ\n- Khối ...	[62492]	161615
1	Nội dung lồng ghép vấn đề bình đẳng giới trong...	[Nội dung lồng ghép vấn đề bình đẳng giới tro...	[151154]	80037
2	Sản phẩm phần mềm có được hưởng ưu đãi về thời...	[Điều 20. Ưu đãi về thời gian miễn thuế, giả...	[75071]	124074
3	Điều kiện để giáo viên trong cơ sở giáo dục mầ...	[Điều kiện được hưởng\nCán bộ quản lý, giáo v...	[225897]	146841
4	Nguyên tắc áp dụng phụ cấp ưu đãi nghề y tế th...	[Nguyên tắc áp dụng\n1. Trường hợp công chức,...	[68365]	6176

# Experimental & Results

02

## Results

	non chunk	chunk	chunk+finetune +rerank BGE
Jinaa-Embedding_8194	44%	47%	57%
Jinaa-Embedding_1024	46%	48%	60%

Kết quả thực nghiệm 1 số phương pháp

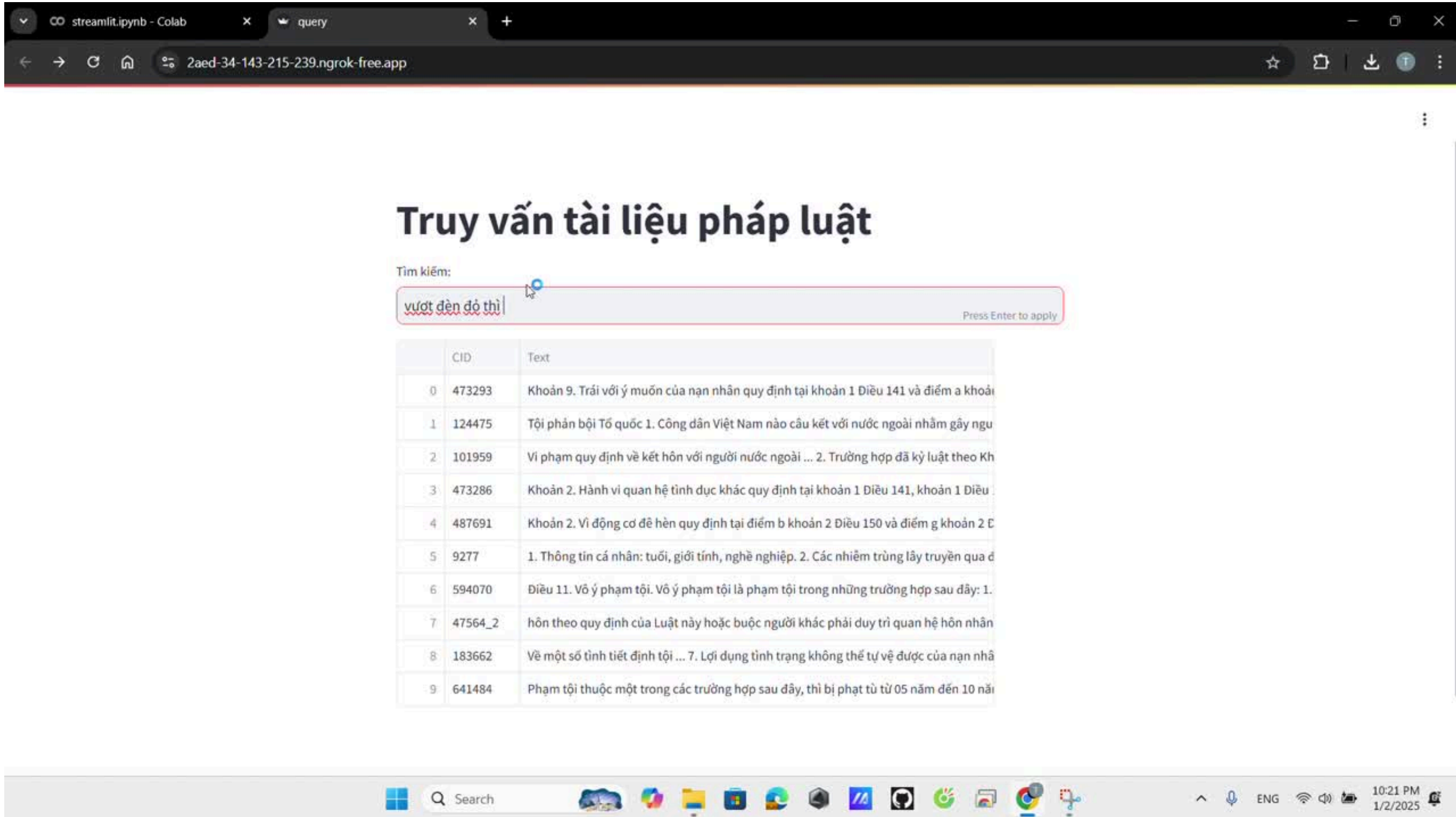
# Conclusion

03

## Conclusion

- Hệ thống đã nâng hiệu suất truy vấn từ 46% lên 60% nhờ sử dụng chunking, fine-tuning Jinaa-Embedding và reranking với BGE.
- Hệ thống đòi hỏi nhiều tài nguyên tính toán và chưa được thử nghiệm đầy đủ trong môi trường tài nguyên hạn chế.
- Chưa khảo sát đủ các trường hợp max sequence length tối ưu cho mô hình
- Chưa sáng tạo trong việc fine-tune model
- Cần chọn model phù hợp hơn
- cần chọn chunk size phù hợp giữa các document để đồng đều và cần tiền xử lý, tổ chức lại documents

# DEMO



streamlit.ipynb - Colab query

2aed-34-143-215-239.ngrok-free.app

## Truy vấn tài liệu pháp luật

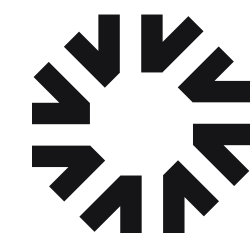
Tìm kiếm:

vượt đèn đỏ thì Press Enter to apply

	CID	Text
0	473293	Khoản 9. Trái với ý muốn của nạn nhân quy định tại khoản 1 Điều 141 và điểm a khoản
1	124475	Tội phản bội Tổ quốc 1. Công dân Việt Nam nào cầu kết với nước ngoài nhằm gây ngu
2	101959	Vi phạm quy định về kết hôn với người nước ngoài ... 2. Trường hợp đã ký luật theo Kh
3	473286	Khoản 2. Hành vi quan hệ tình dục khác quy định tại khoản 1 Điều 141, khoản 1 Điều .
4	487691	Khoản 2. Vi động cơ đề hèn quy định tại điểm b khoản 2 Điều 150 và điểm g khoản 2 Đ
5	9277	1. Thông tin cá nhân: tuổi, giới tính, nghề nghiệp. 2. Các nhiễm trùng lây truyền qua đ
6	594070	Điều 11. Vô ý phạm tội. Vô ý phạm tội là phạm tội trong những trường hợp sau đây: 1.
7	47564_2	hôn theo quy định của Luật này hoặc buộc người khác phải duy trì quan hệ hôn nhân
8	183662	Về một số tình tiết định tội ... 7. Lợi dụng tình trạng không thể tự vệ được của nạn nhâ
9	641484	Phạm tội thuộc một trong các trường hợp sau đây, thì bị phạt tù từ 05 năm đến 10 năm

Search

ENG 10:21 PM 1/2/2025



CS336

***Thank You***