RAG System for Efficient Access to University Admissions Information

Riro Jin , Jaehyeok Shin

Supervisor: Prof. Misuk Kim

School of Data Science, Hanyang University

Research Motivations

Hallucination

Eliminating hallucinations and ensuring consistent answers still remains difficult.

Limitations of existing chatbots

Most admission chatbots focus on business-oriented information.

↕

Applicants actually need precise guideline details.

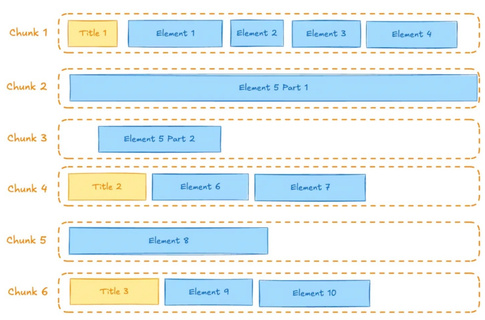
To bridge the gap between what companies provide and what users genuinely need, this project proposes a "RAG-based admissions guideline information system".

Data Preprocessing

- We performed text-based preprocessing using *pdfplumber* and *Tesseract OCR*.

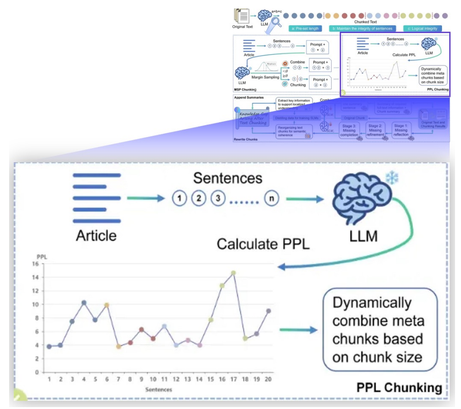
Related Works

Financial Report Chunking for Effective Retrieval Augmented Generation (2024)



[element-based chunking]

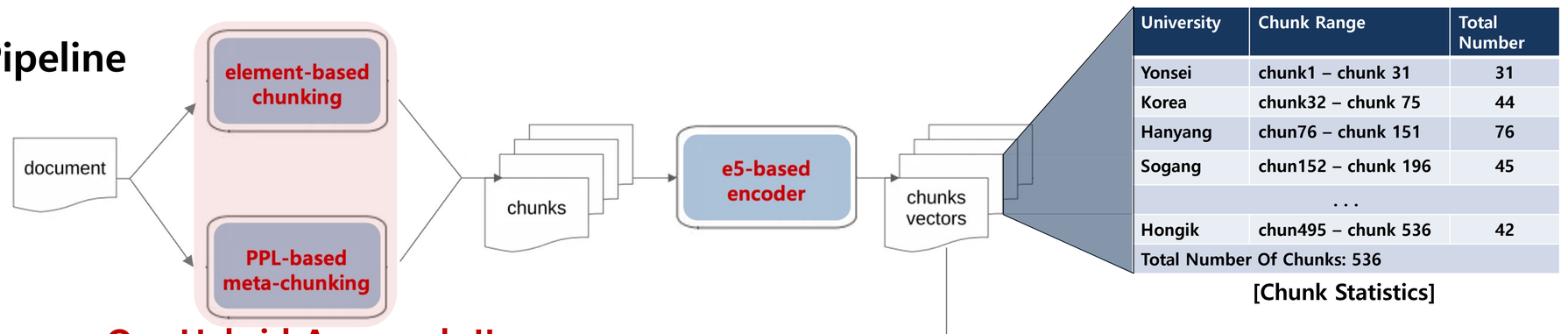
Meta-Chunking: Learning Text Segmentation and Semantic Completion via Logical Perception (2024)



[PPL-based meta-chunking]

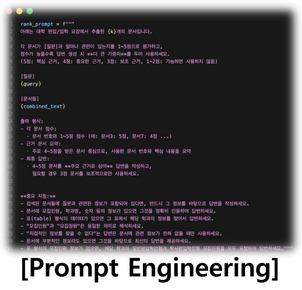
Model Development

Model Pipeline

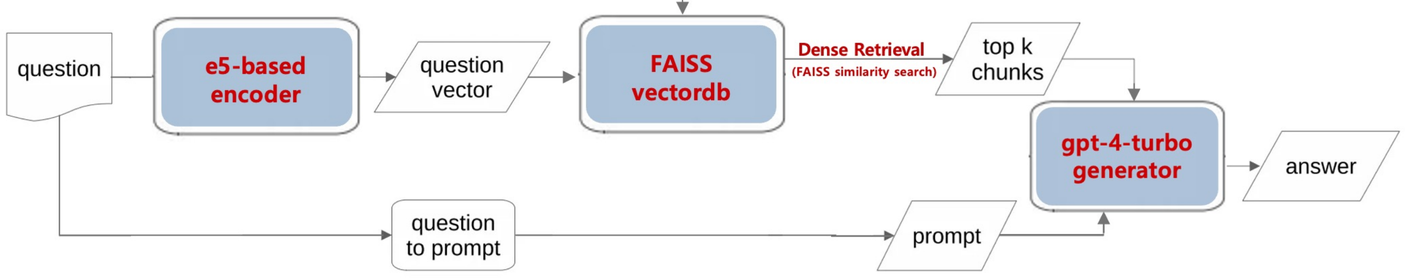


University	Chunk Range	Total Number
Yonsei	chunk1 – chunk 31	31
Korea	chunk32 – chunk 75	44
Hanyang	chun76 – chunk 151	76
Sogang	chun152 – chunk 196	45
. . .		
Hongik	chun495 – chun 536	42
Total Number Of Chunks: 536		

[Chunk Statistics]



[Prompt Engineering]



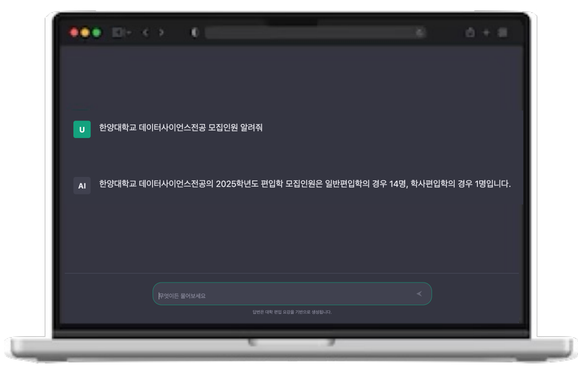
Ablation Study

Model architecture	F1-score
Fixed-length Chunking	0.5805
Element-based Chunking	0.6062
PPL-based Meta Chunking	0.6501
Element-based Chunking & PPL-based Meta Chunking	0.7099

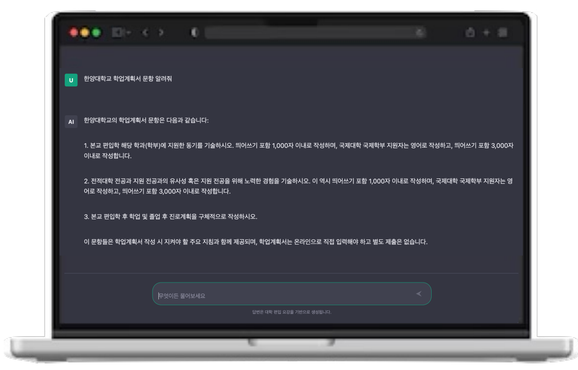
12% ↑

- The hybrid approach that combines Element-based Chunking and PPL-based Meta Chunking achieved the highest performance (F1 0.7099).

Use Cases



[Case1: Extracting the table by element-based chunking]



[Case2: Preserving context in plain text by PPL-based meta chunking]

Project Progress & Future Plan

Mar – Apr – May – Jun – Jul – Aug – Sep – Oct – Nov – Dec – Future

- Mar–Apr: PDF Text Preprocessing
- Apr–May: Fixed-length Chunking
- May–Jun: Related Works Review
- Jun–Sep: Hybrid Chunking Development
- Sep–Nov: Pipeline Construction
- Nov–Dec: LLM Integration & Prompt Engineering
- Future Plan: Developing a Personalized University Recommendation System