Data Parsing

Document Structuring

Peter

데이터팀

Who is Speaker?

Hwang Jaesung (Peter)

AI Research Engineer

Braincrew Inc. Data 팀 Leader

다양한 Data Processing

- 23.09~25.08
  - Bio Domain LLM optimization
- 25.06~25.11
  - BrainCrew - RAG Project
- 25.11~
  - BrainCrew - Data Team Leader

Linked-In

Too Much Information

Data Team 하는 일

[Job]

- Document Processing – AI Comprehension
  - Experiments
  - Pipelines
- Parsing AGENT
- Streaming Data Processing with AGENT
- Data Pipelines
  - For Embedding Models
  - ...
- (Synthetic Data, Ontologies, ...)

Too Much Information

Data Team 하는 일

[Job]

- Document Processing – AI Comprehension
  - Experiments
  - Pipelines
- Parsing AGENT
- Streaming Data Processing with AGENT
- Data Pipelines
  - For Embedding Models
  - ...
- (Synthetic Data, Ontologies, ...)

"단순히 읽는 게 아니라
AI가 이해할 수 있게 하기"

"기본적이고 중요하지만 간과하는"

CONTENTS

PART 01

Structuring의 중요성

"LLM이 알아서 다 해준다?"

01

RAG 최근 동향

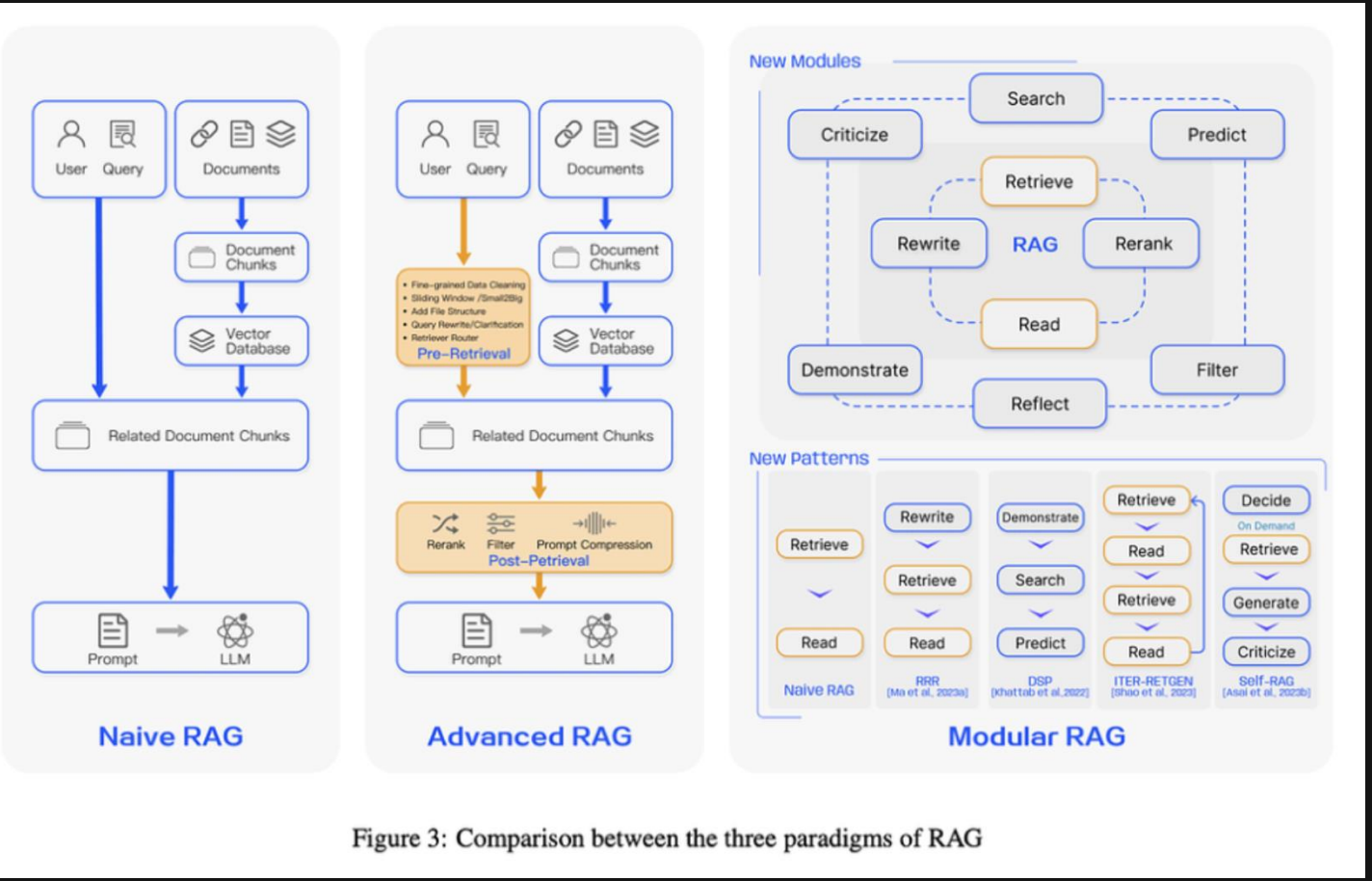
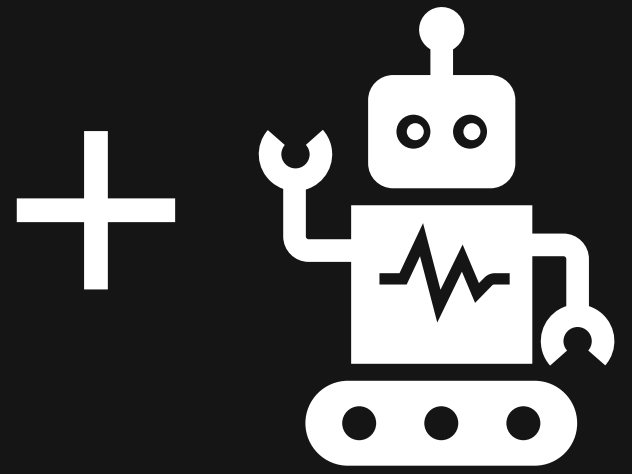


Figure 3: Comparison between the three paradigms of RAG



RAG 최근 동향

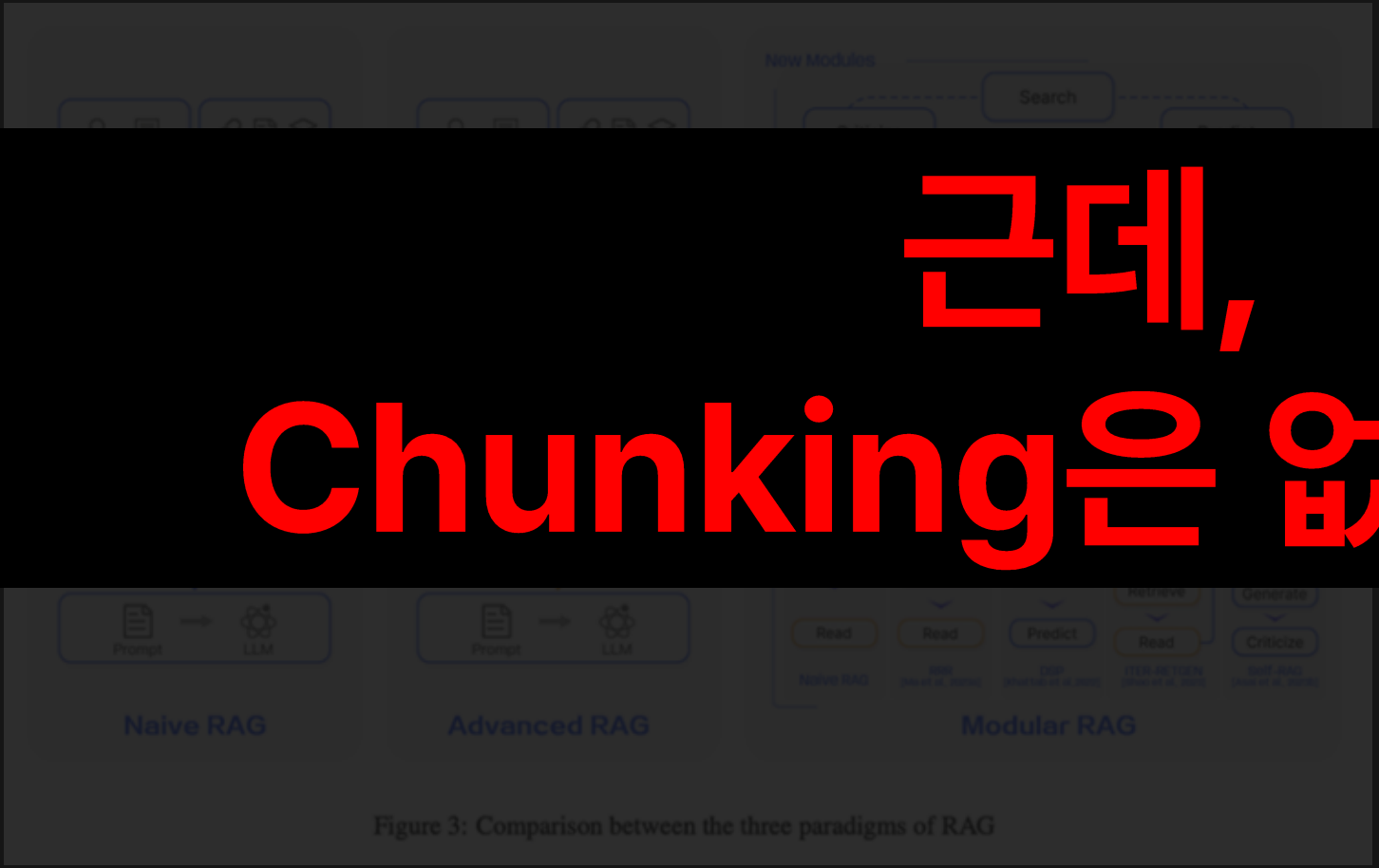


Figure 3: Comparison between the three paradigms of RAG

근데,
Chunking은 없잖아 ...

기업 문서에 대한 RAG

단일 문서라면?

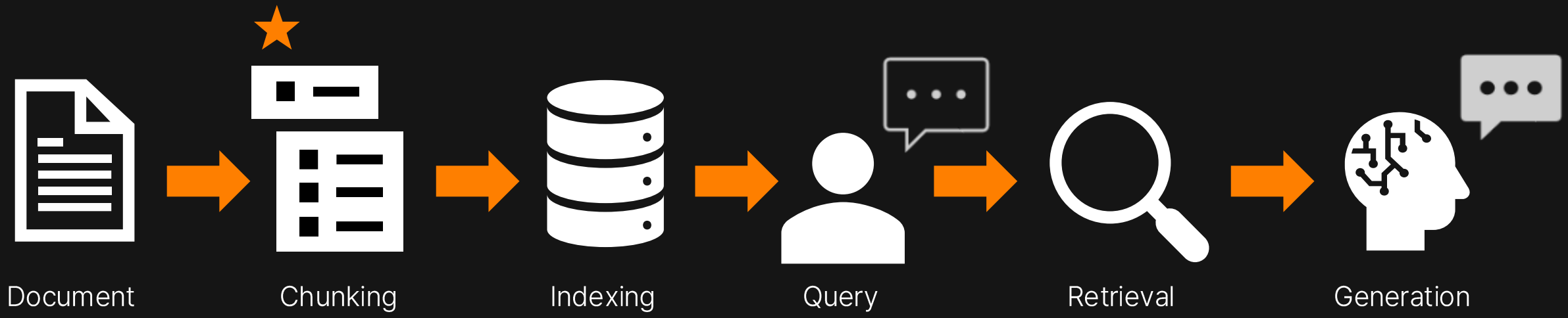
단일 문서

기업 문서에 대한 RAG

기업 문서라면?

기업 문서에 대한 RAG

결론: 방대한 기업 문서에서 가장 중요한 건 Chunking!!



돌고 돌아 Chunking

검색을 잘 하려면 문서의 구조정보를 활용하여 의미 단위로 구분하는 것부터

Fixed-size Chunking

LLM based Chunking

Recursive Chunking

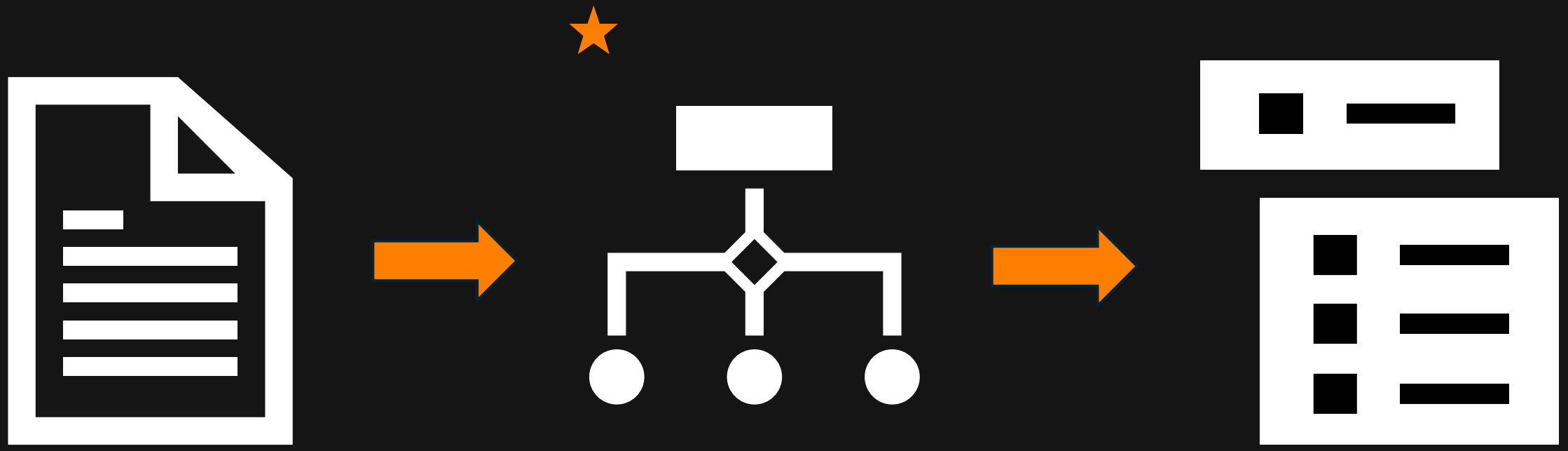
Late Chunking

Semantic Chunking

결국에 목표는 문서의 구조 정보와 의미 정보를 잘 활용하여 의미 단위로 잘 자르는 것

Document Structuring

그러면 그 Chunking을 잘하려면?



Document

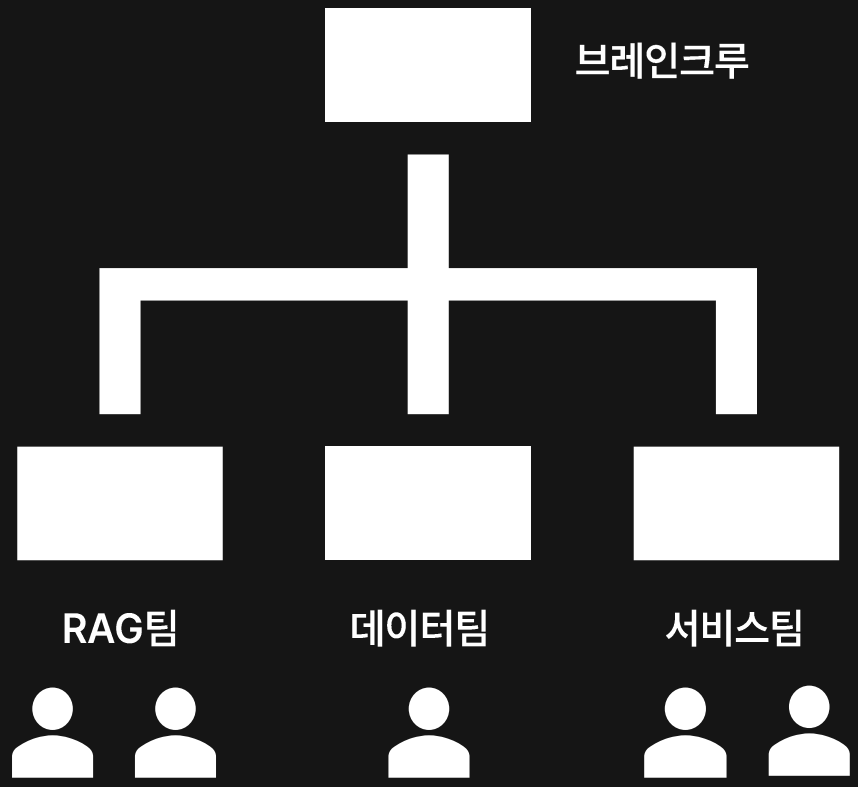
Structuring

Chunking

Structuring이 더 우선시 되어야 함

Document Structuring

왜?



Document Structuring이 중요한 이유

조직도처럼 문서는 구조라는 뼈와 의미라는 살로 이루어짐

Towards Robust Diabetes-Specialized LLMs: Reflection- and Curriculum-based Instruction Tuning Across Diverse Tasks

Jaesung Hwang, MSc[1] and Donghyeon Park, PhD[2]

[1]*Department of Software, Sejong University, Seoul 05006, Republic of Korea.*
[2]*Department of Artificial Intelligence & Data Science, Sejong University, Seoul 05006, Republic of Korea.*

Abstract

Objective: Diabetes is a globally prevalent and complex chronic condition that requires continuous interpretation of glycemic trends, personalized lifestyle interventions, and sustained patient education. While large language models (LLMs) have shown promise in various biomedical applications, existing general-purpose and biomedical LLMs often underperform in diabetes-specific reasoning and instruction-following tasks. To address this limitation, we present a diabetes-specialized LLM fine-tuned on a curated instruction dataset that reflects the nuances of diabetes management.

Methods: Our approach integrates two reflection-based replacement metrics—Instruction-Following Difficulty (IFD) and reversed-IFD (r-IFD)—to evaluate the model's receptiveness to instructions and to filter and substitute suboptimal prompts. In addition, we apply a curriculum instruction tuning strategy that sequences instructions from easier to more challenging based on model sensitivity, mitigating catastrophic forgetting while promoting stable acquisition of domain expertise.

Results: We evaluate the model across multiple diabetes-related task categories, including question answering, natural language inference, information extraction, summarization, generation, and diabetes-friendly dietary recommendation. Our model demonstrates superior performance in accuracy compared to strong biomedical baselines, and generates meal plans that are 0.66% more suitable than those produced by GPT-4, as measured by the Diet Quality Index-International (DQI-I)—an international metric for nutritional quality. Furthermore, simulation-based evaluations using simglucose reveal that our model's meal plans lead to more favorable glycemic outcomes compared to GPT-4.

Conclusion: These results highlight the potential of domain-specific instruction tuning to transform general-purpose LLMs into expert-level assistants for diabetes management.

Key words: Diabetes Management, Reflection-based Replacement, Instruction Following Difficulty, Reversed-Instruction Following Difficulty, Curriculum Instruction Tuning, diabetes-friendly Nutrition, Glycemic Simulation

Introduction

Diabetes is one of the most prevalent and complex chronic diseases globally, currently affecting more than 500 million individuals and contributing significantly to the global burden of non-communicable diseases. According to a comprehensive pooled analysis of 1108 population-representative studies including over 141 million participants, the worldwide prevalence of diabetes has risen continuously from 1990 to 2022, and it is projected to escalate even further in the next decades[25]. Managing diabetes effectively requires not only pharmacological treatment but also ongoing interpretation of blood glucose fluctuations, dietary adaptations, personalized lifestyle interventions, and continuous patient education. As the demand for scalable, patient-centered care rises, the healthcare community increasingly seeks robust AI tools, such as large language models (LLMs), that can support clinical decision-making and self-management guidance.

LLMs have recently gained traction in the biomedical domain due to their capability to generate human-like responses and synthesize complex information. In the context of diabetes management, large language models (LLMs) have been leveraged across a diverse set of tasks tailored to patient needs. For instance, integrated systems combining image-based data with LLM-generated responses have been shown to enhance interpretability in primary management settings[11]. Retrieval-augmented generation (RAG) architectures have been employed to support diabetes education by grounding LLM outputs in trusted reference sources, thereby improving factual reliability[17]. In another example, large language models have been applied to analyze CGM data directly and summarize trends over time for patients and clinicians[1]. But, there is no practical diabetes management LLM that answers questions about diabetes and even provides diet recommendations.

Also, these promising directions still face critical limitations. Biomedical LLMs such as BioBERT[5] and BioGPT[12] have demonstrated high performance on general biomedical QA benchmarks but often underperform on reasoning tasks related to diabetes management, which require clinical nuance, contextual understanding, and personalized judgment. This limitation was further substantiated in a recent prospective study, where even advanced biomedical LLMs failed to achieve satisfactory accuracy and alignment when evaluated on Multiple diabetes-related task. Similarly, general-purpose LLMs like ChatGPT and Bard continue to struggle with endocrinology-related queries, frequently providing vague or partially incorrect answers that lack alignment with clinical guidelines[4]. These limitations underscore the need for models that are not only trained on diabetes-specific data but also optimized for diabetes-specific reasoning and instruction-following in diabetes management, including accurate answers to diabetes-related questions and diabetes-friendly dietary recommendations.

To address this gap, we propose a diabetes-specialized large language model trained on a diabetes-specific instruction dataset designed to reflect the nuances of diabetes management. The dataset includes a diverse set of prompts covering glycemic trend interpretation, insulin dose adjustment, nutritional guidance, symptom analysis, and patient counseling. Importantly, the model is trained not merely to recall factual information but to engage in higher-order reasoning that emulates the decision-making patterns of healthcare professionals. To support this goal, we go beyond merely

=

```
Document
├── Title
│   └── Towards Robust Diabetes-Specialized LLMs:
│       Reflection- and Curriculum-based Instruction Tuning Across Diverse Tasks
├── Authors
│   ├── Jaesung Hwang, MSc
│   └── Donghyeon Park, PhD
├── Affiliations
│   ├── Department of Software, Sejong University
│   └── Department of Artificial Intelligence & Data Science, Sejong University
├── Abstract
│   ├── Objective
│   ├── Methods
│   ├── Results
│   └── Conclusion
├── Keywords
│   └── Diabetes Management, Reflection-based Replacement,
│       Instruction Following Difficulty, Curriculum Instruction Tuning,
│       Glycemic Simulation
└── Section 1: Introduction
```

이러한 구조화 정보를 활용하면 더 의미적

Document Structuring이 중요한 이유

검색 관점 이점

설명

문서는 원래 구조 정보를 가지고 있음

- Parser만 사용하면 구조화가 어려움

검색 시에는 작은 Chunk, 답변 생성 시에는 큰 Chunk가 유리함

- 너무 큰 Chunk는 의미 희석, 너무 작은 Chunk는 답변 생성 정보 부족

검색 관점 이점

문맥 소실 방지

- 하위 계층을 검색하더라도 상위 계층을 함께 전달해줄 수 있음

다양한 질문 유형에 대한 대응

- “A제품의 나사 규격은?”, “A제품의 전반적인 유지보수 방법은?” 과 같은 구체적 질문과 포괄적 질문에 모두 대응 가능

Document Structuring이 중요한 이유

데이터 관점

정보 이론 관점

정보 밀도의 최적화

- 질문 – 고밀도, 문서 – 저밀도
- 문서의 하위 계층 – 고밀도, 문서의 상위 계층 – 저밀도
- -> 검색은 고밀도 공간에서 하고, 생성은 저밀도 공간에서 수행 가능

벡터 공간에서의 희석 문제

- 하나의 Chunk에 A, B, C 주제가 섞여 있다면 평균화 오류 발생
- -> 각 Vector는 한 가지 주제에만 집중된 분포를 가짐

검색 점수의 Long Tail 방지

- Top-K를 가져와도 핵심 정보의 일부만 포함될 확률이 높음
- -> 구조화를 통해 Parent-Child 구조로 해결 가능

Document Structuring이 중요한 이유

이론 말고, 실제로는?

사례 공유

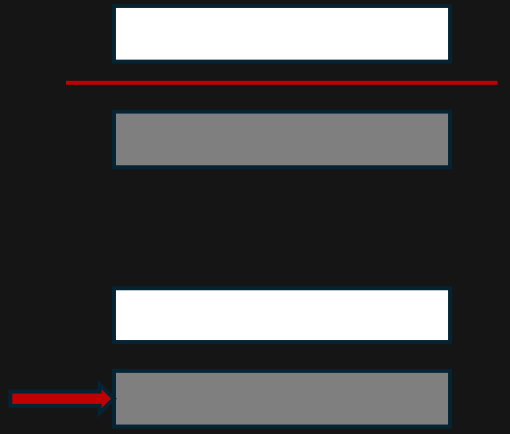
Chunking 기법 평가

건설 도메인 RAG 프로젝트 특정 공종

- Structuring 기법으로 Recall 85%

MoC: Mixtures of Text Chunking Learners for RAG system

- Chunk 성능 평가 방법 제안
- Boundary Clarify (BC)
  - 문서에서 주제가 바뀌는 지점을 Chunking이 잘 끊어주고 있는지
- Chunk Stickiness (CS)
  - 하나의 Chunk 안에 문장들이 같은 이야기를 하고 있는지



Document Structuring이 중요한 이유

검증 (feat. Semantic Distance)

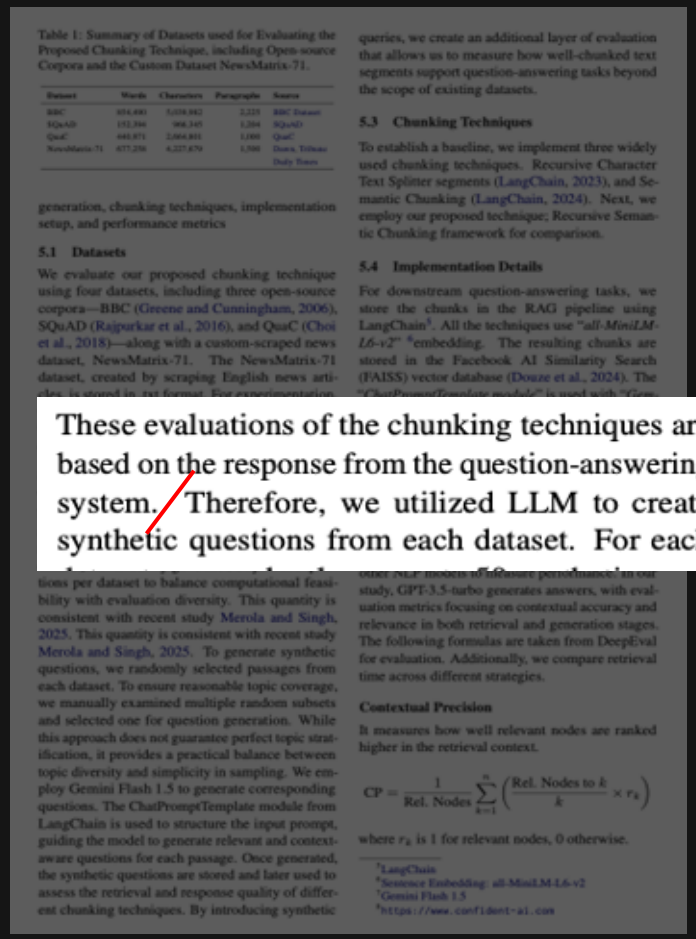


Table 1: Summary of Datasets used for Evaluating the Proposed Chunking Technique, including Open-source Corpora and the Custom Dataset NewsMatrix-71.

Dataset	Words	Characters	Paragraphs	Source
BBC	854,480	5,038,882	2,225	BBC Dataset
SQuAD	152,306	908,345	1,204	SQuAD
QuaC	641,871	2,964,801	1,000	QuaC
NewsMatrix-71	677,258	4,227,879	1,500	Dawn, Tribune, Daily Times

generation, chunking techniques, implementation setup, and performance metrics

5.1 Datasets

We evaluate our proposed chunking technique using four datasets, including three open-source corpora—BBC (Greene and Cunningham, 2006), SQuAD (Rajpurkar et al., 2016), and QuaC (Choi et al., 2018)—along with a custom-scraped news dataset, NewsMatrix-71. The NewsMatrix-71 dataset, created by scraping English news arti-

These evaluations of the chunking techniques are based on the response from the question-answering system. Therefore, we utilized LLM to create synthetic questions from each dataset. For each

tions per dataset to balance computational feasibility with evaluation diversity. This quantity is consistent with recent study Merola and Singh, 2025. This quantity is consistent with recent study Merola and Singh, 2025. To generate synthetic questions, we randomly selected passages from each dataset. To ensure reasonable topic coverage, we manually examined multiple random subsets and selected one for question generation. While this approach does not guarantee perfect topic stratification, it provides a practical balance between topic diversity and simplicity in sampling. We employ Gemini Flash 1.5 to generate corresponding questions. The ChatPromptTemplate module from LangChain is used to structure the input prompt, guiding the model to generate relevant and context-aware questions for each passage. Once generated, the synthetic questions are stored and later used to assess the retrieval and response quality of different chunking techniques. By introducing synthetic

queries, we create an additional layer of evaluation that allows us to measure how well-chunked text segments support question-answering tasks beyond the scope of existing datasets.

5.3 Chunking Techniques

To establish a baseline, we implement three widely used chunking techniques. Recursive Character Text Splitter segments (LangChain, 2023), and Semantic Chunking (LangChain, 2024). Next, we employ our proposed technique; Recursive Semantic Chunking framework for comparison.

5.4 Implementation Details

For downstream question-answering tasks, we store the chunks in the RAG pipeline using LangChain[3]. All the techniques use "*all-MiniLM-L6-v2*" [4]embedding. The resulting chunks are stored in the Facebook AI Similarity Search (FAISS) vector database (Douze et al., 2024). The

study, GPT-3.5-turbo generates answers, with evaluation metrics focusing on contextual accuracy and relevance in both retrieval and generation stages. The following formulas are taken from DeepEval for evaluation. Additionally, we compare retrieval time across different strategies.

Contextual Precision

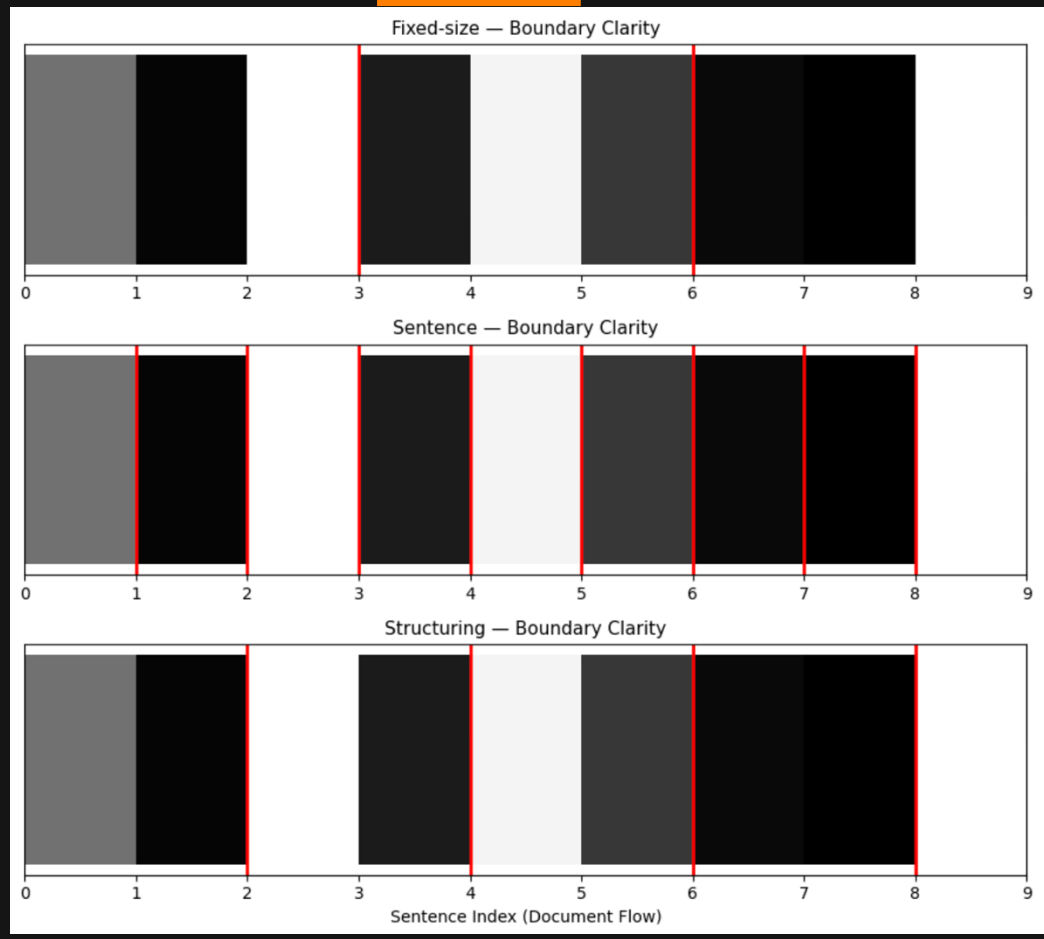
It measures how well relevant nodes are ranked higher in the retrieval context.

$$CP = \frac{1}{\text{Rel. Nodes}} \sum_{k=1}^{n} \left(\frac{\text{Rel. Nodes to } k}{k} \times r_k \right)$$

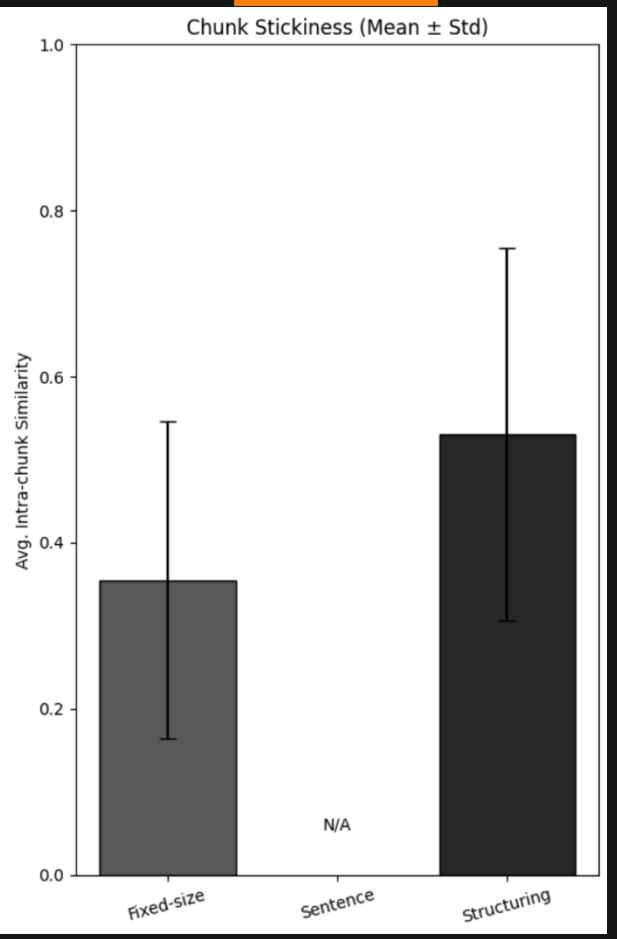
where r_k is 1 for relevant nodes, 0 otherwise.

[3]LangChain
[4]Sentence Embedding: all-MiniLM-L6-v2
[5]Gemini Flash 1.5
[6]https://www.confident-ai.com

BC



CS



Reference -MoC: Mixtures of Text Chunking Learners for Retrieval-Augmented Generation System

PART 02

Regex로 구조 잡기

"사람이 LLM이 되는 방법"

02

Regex로 규칙 찾기..?

Table 1: Summary of Datasets used for Evaluating the Proposed Chunking Technique, including Open-source Corpora and the Custom Dataset NewsMatrix-71.

Dataset	Words	Characters	Paragraphs	Source
BBC	854,490	5,039,982	2,225	BBC Dataset
SQuAD	152,394	966,345	1,204	SQuAD
QuaC	440,971	2,664,801	1,000	QuaC
NewsMatrix-71	677,258	4,227,679	1,500	Dawn, Tribune Daily Times

generation, chunking techniques, implementation setup, and performance metrics

5.1 Datasets

We evaluate our proposed chunking technique using four datasets, including three open-source corpora—BBC (Greene and Cunningham, 2006), SQuAD (Rajpurkar et al., 2016), and QuaC (Choi et al., 2018)—along with a custom-scraped news dataset, NewsMatrix-71. The NewsMatrix-71 dataset, created by scraping English news articles, is stored in .txt format. For experimentation, we use a 1,500-article subset containing 677,258 words and 4,227,679 characters. A summary of all datasets is provided in Table 1.

5.2 Synthetic Question Generation

These evaluations of the chunking techniques are based on the response from the question-answering system. Therefore, we utilized LLM to create synthetic questions from each dataset. For each dataset, we randomly generate 50 synthetic questions per dataset to balance computational feasibility with evaluation diversity. This quantity is consistent with recent study Merola and Singh, 2025. This quantity is consistent with recent study Merola and Singh, 2025. To generate synthetic questions, we randomly selected passages from each dataset. To ensure reasonable topic coverage, we manually examined multiple random subsets and selected one for question generation. While this approach does not guarantee perfect topic stratification, it provides a practical balance between topic diversity and simplicity in sampling. We employ Gemini Flash 1.5 to generate corresponding questions. The ChatPromptTemplate module from LangChain is used to structure the input prompt, guiding the model to generate relevant and context-aware questions for each passage. Once generated, the synthetic questions are stored and later used to assess the retrieval and response quality of different chunking techniques. By introducing synthetic queries, we create an additional layer of evaluation that allows us to measure how well-chunked text segments support question-answering tasks beyond the scope of existing datasets.

5.3 Chunking Techniques

To establish a baseline, we implement three widely used chunking techniques. Recursive Character Text Splitter segments (LangChain, 2023), and Semantic Chunking (LangChain, 2024). Next, we employ our proposed technique; Recursive Semantic Chunking framework for comparison.

5.4 Implementation Details

For downstream question-answering tasks, we store the chunks in the RAG pipeline using LangChain[5]. All the techniques use "*all-MiniLM-L6-v2*" [6]embedding. The resulting chunks are stored in the Facebook AI Similarity Search (FAISS) vector database (Douze et al., 2024). The "*ChatPromptTemplate module*" is used with "*Gemini Flash 1.5*" [7], a state-of-the-art Large Language Model optimized for contextual reasoning.

5.5 Evaluation metrics

We assess chunking techniques by integrating them into the RAG pipeline for a question-answering task. For evaluation, we use DeepEval by Confident AI [8], an open-source framework designed for LLM evaluation. DeepEval leverages LLMs and other NLP models to measure performance. In our study, GPT-3.5-turbo generates answers, with evaluation metrics focusing on contextual accuracy and relevance in both retrieval and generation stages. The following formulas are taken from DeepEval for evaluation. Additionally, we compare retrieval time across different strategies.

Contextual Precision

It measures how well relevant nodes are ranked higher in the retrieval context.

$$\text{CP} = \frac{1}{\text{Rel. Nodes}} \sum_{k=1}^{n} \left(\frac{\text{Rel. Nodes to } k}{k} \times r_k \right)$$

where r_k is 1 for relevant nodes, 0 otherwise.

[5] LangChain
[6] Sentence Embedding: all-MiniLM-L6-v2
[7] Gemini Flash 1.5
[8] https://www.confident-ai.com

어떻게?

Regex로 규칙 찾기..?

Heading의 규칙을 찾아 구조화

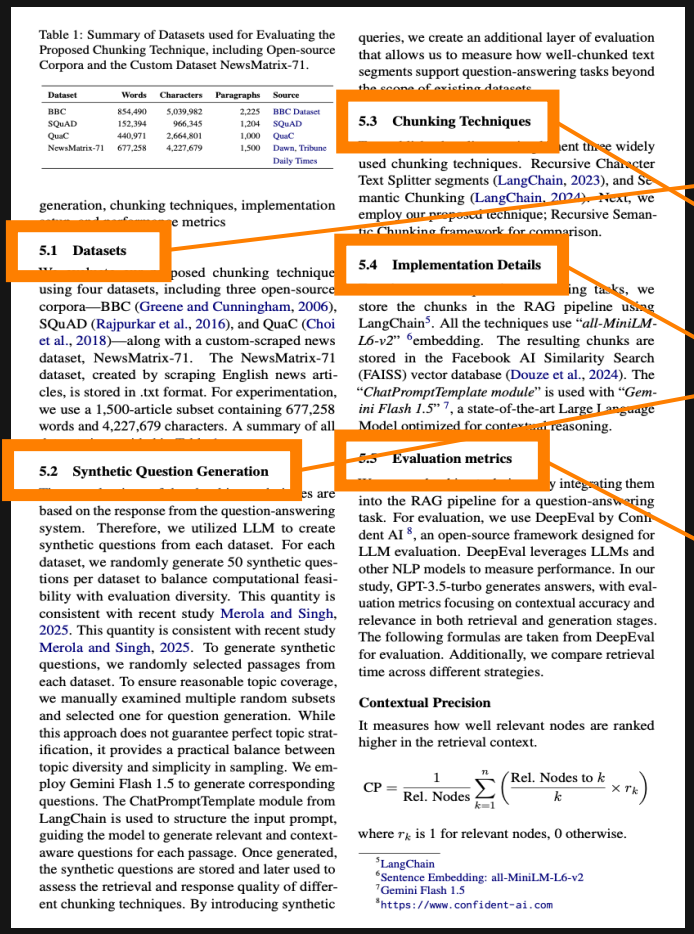


Table 1: Summary of Datasets used for Evaluating the Proposed Chunking Technique, including Open-source Corpora and the Custom Dataset NewsMatrix-71.

Dataset	Words	Characters	Paragraphs	Source
BBC	854,490	5,039,982	2,225	BBC Dataset
SQuAD	152,394	966,345	1,204	SQuAD
QuaC	440,971	2,664,801	1,000	QuaC
NewsMatrix-71	677,258	4,227,679	1,500	Dawn, Tribune Daily Times

generation, chunking techniques, implementation ... metrics

5.1 Datasets

...posed chunking technique ... using four datasets, including three open-source corpora—BBC (Greene and Cunningham, 2006), SQuAD (Rajpurkar et al., 2016), and QuaC (Choi et al., 2018)—along with a custom-scraped news dataset, NewsMatrix-71. The NewsMatrix-71 dataset, created by scraping English news articles, is stored in .txt format. For experimentation, we use a 1,500-article subset containing 677,258 words and 4,227,679 characters. A summary of all ...

5.2 Synthetic Question Generation

...es are based on the response from the question-answering system. Therefore, we utilized LLM to create synthetic questions from each dataset. For each dataset, we randomly generate 50 synthetic questions per dataset to balance computational feasibility with evaluation diversity. This quantity is consistent with recent study Merola and Singh, 2025. This quantity is consistent with recent study Merola and Singh, 2025. To generate synthetic questions, we randomly selected passages from each dataset. To ensure reasonable topic coverage, we manually examined multiple random subsets and selected one for question generation. While this approach does not guarantee perfect topic stratification, it provides a practical balance between topic diversity and simplicity in sampling. We employ Gemini Flash 1.5 to generate corresponding questions. The ChatPromptTemplate module from LangChain is used to structure the input prompt, guiding the model to generate relevant and context-aware questions for each passage. Once generated, the synthetic questions are stored and later used to assess the retrieval and response quality of different chunking techniques. By introducing synthetic queries, we create an additional layer of evaluation that allows us to measure how well-chunked text segments support question-answering tasks beyond the scope of existing datasets.

5.3 Chunking Techniques

...ent three widely used chunking techniques. Recursive Character Text Splitter segments (LangChain, 2023), and Semantic Chunking (LangChain, 2024). Next, we employ our proposed technique; Recursive Semantic Chunking framework for comparison.

5.4 Implementation Details

...ing tasks, we store the chunks in the RAG pipeline using LangChain[5]. All the techniques use "*all-MiniLM-L6-v2*" [6]embedding. The resulting chunks are stored in the Facebook AI Similarity Search (FAISS) vector database (Douze et al., 2024). The "*ChatPromptTemplate module*" is used with "*Gemini Flash 1.5*" [7], a state-of-the-art Large Language Model optimized for contextual reasoning.

5.5 Evaluation metrics

...y integrating them into the RAG pipeline for a question-answering task. For evaluation, we use DeepEval by Confident AI [8], an open-source framework designed for LLM evaluation. DeepEval leverages LLMs and other NLP models to measure performance. In our study, GPT-3.5-turbo generates answers, with evaluation metrics focusing on contextual accuracy and relevance in both retrieval and generation stages. The following formulas are taken from DeepEval for evaluation. Additionally, we compare retrieval time across different strategies.

Contextual Precision

It measures how well relevant nodes are ranked higher in the retrieval context.

$$\text{CP} = \frac{1}{\text{Rel. Nodes}} \sum_{k=1}^{n} \left(\frac{\text{Rel. Nodes to } k}{k} \times r_k \right)$$

where r_k is 1 for relevant nodes, 0 otherwise.

[5]LangChain
[6]Sentence Embedding: all-MiniLM-L6-v2
[7]Gemini Flash 1.5
[8]https://www.confident-ai.com

5, 5.1, 5.2,.. 등 prefix가 있으면 효과적

5.1 Datasets

5.2 Synthetic Question Generation

5.3 Chunking Techniques

5.4 Implementation Details

5.5 Evaluation metrics

Regex로 규칙 찾기..?

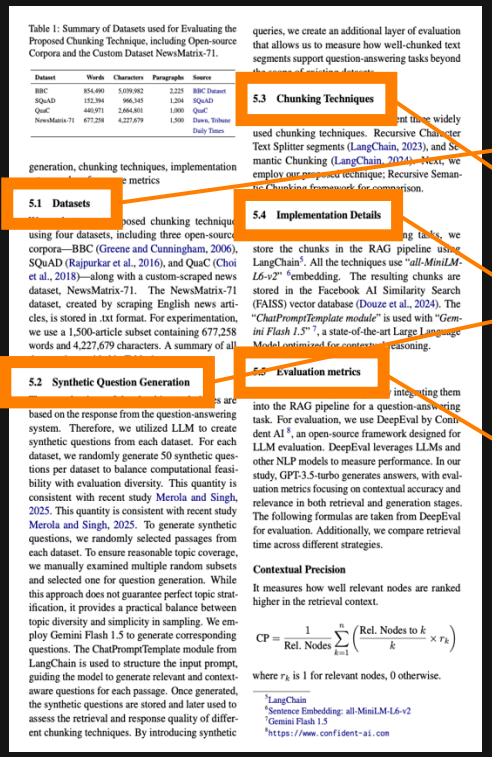


Table 1: Summary of Datasets used for Evaluating the Proposed Chunking Technique, including Open-source Corpora and the Custom Dataset NewsMatrix-71.

Dataset	Words	Characters	Paragraphs	Source
BBC	854,490	5,039,982	2,225	BBC Dataset
SQuAD	152,394	966,345	1,204	SQuAD
QuaC	440,971	2,664,801	1,000	QuaC
NewsMatrix-71	677,258	4,227,679	1,500	Dawn, Tribune Daily Times

generation, chunking techniques, implementation metrics

5.1 Datasets

using four datasets, including three open-source corpora—BBC (Greene and Cunningham, 2006), SQuAD (Rajpurkar et al., 2016), and QuaC (Choi et al., 2018)—along with a custom-scraped news dataset, NewsMatrix-71. The NewsMatrix-71 dataset, created by scraping English news articles, is stored in .txt format. For experimentation, we use a 1,500-article subset containing 677,258 words and 4,227,679 characters. A summary of all

5.2 Synthetic Question Generation

based on the response from the question-answering system. Therefore, we utilized LLM to create synthetic questions from each dataset. For each dataset, we randomly generate 50 synthetic questions per dataset to balance computational feasibility with evaluation diversity. This quantity is consistent with recent study Merola and Singh, 2025. This quantity is consistent with recent study Merola and Singh, 2025. To generate synthetic questions, we randomly selected passages from each dataset. To ensure reasonable topic coverage, we manually examined multiple random subsets and selected one for question generation. While this approach does not guarantee perfect topic stratification, it provides a practical balance between topic diversity and simplicity in sampling. We employ Gemini Flash 1.5 to generate corresponding questions. The ChatPromptTemplate module from LangChain is used to structure the input prompt, guiding the model to generate relevant and context-aware questions for each passage. Once generated, the synthetic questions are stored and later used to assess the retrieval and response quality of different chunking techniques. By introducing synthetic queries, we create an additional layer of evaluation that allows us to measure how well-chunked text segments support question-answering tasks beyond

5.3 Chunking Techniques

used chunking techniques. Recursive Character Text Splitter segments (LangChain, 2023), and Semantic Chunking (LangChain, 2024). Next, we employ our proposed technique; Recursive Semantic Chunking framework for comparison.

5.4 Implementation Details

store the chunks in the RAG pipeline using LangChain[5]. All the techniques use *"all-MiniLM-L6-v2"* [6]embedding. The resulting chunks are stored in the Facebook AI Similarity Search (FAISS) vector database (Douze et al., 2024). The *"ChatPromptTemplate module"* is used with *"Gemini Flash 1.5"* [7], a state-of-the-art Large Language Model optimized for contextual reasoning.

5.5 Evaluation metrics

integrating them into the RAG pipeline for a question-answering task. For evaluation, we use DeepEval by Confident AI [8], an open-source framework designed for LLM evaluation. DeepEval leverages LLMs and other NLP models to measure performance. In our study, GPT-3.5-turbo generates answers, with evaluation metrics focusing on contextual accuracy and relevance in both retrieval and generation stages. The following formulas are taken from DeepEval for evaluation. Additionally, we compare retrieval time across different strategies.

Contextual Precision

It measures how well relevant nodes are ranked higher in the retrieval context.

$$\text{CP} = \frac{1}{\text{Rel. Nodes}} \sum_{k=1}^{n} \left(\frac{\text{Rel. Nodes to } k}{k} \times r_k \right)$$

where r_k is 1 for relevant nodes, 0 otherwise.

[5]LangChain
[6]Sentence Embedding: all-MiniLM-L6-v2
[7]Gemini Flash 1.5
[8]https://www.confident-ai.com

5.1 Datasets

5.2 Synthetic Question Generation

5.3 Chunking Techniques

5.4 Implementation Details

5.5 Evaluation metrics

5.5 Evaluation metrics

We assess chunking techniques by integrating them into the RAG pipeline for a question-answering task. For evaluation, we use DeepEval by Confident AI [8], an open-source framework designed for LLM evaluation. DeepEval leverages LLMs and other NLP models to measure performance. In our study, GPT-3.5-turbo generates answers, with evaluation metrics focusing on contextual accuracy and relevance in both retrieval and generation stages. The following formulas are taken from DeepEval for evaluation. Additionally, we compare retrieval time across different strategies.

의도치 않은 분할 발생

Regex로 규칙 찾기..?

한계점 존재

사람이 직접 LLM이 되어야 함

- 문서 하나하나 특징을 사람이 직접 파악해야 함

무한 예외 상황

- Heading에서 잡은 pattern이 본문 내에도 존재

일관성 이슈

- 규칙이 아무리 촘촘해도 문서마다 서로 다름

구조적 맥락 상실

- 계층적이라는 특성을 가지는 문서의 특성을 살리지 못함

Regex가 잘 되려면

- 문서에 목차가 있으면 잘 될 수 있음

PART 03

LLM으로 구조 추론하기

"효과적이지만 비싸"

03

LLM이 해주지 않을까?

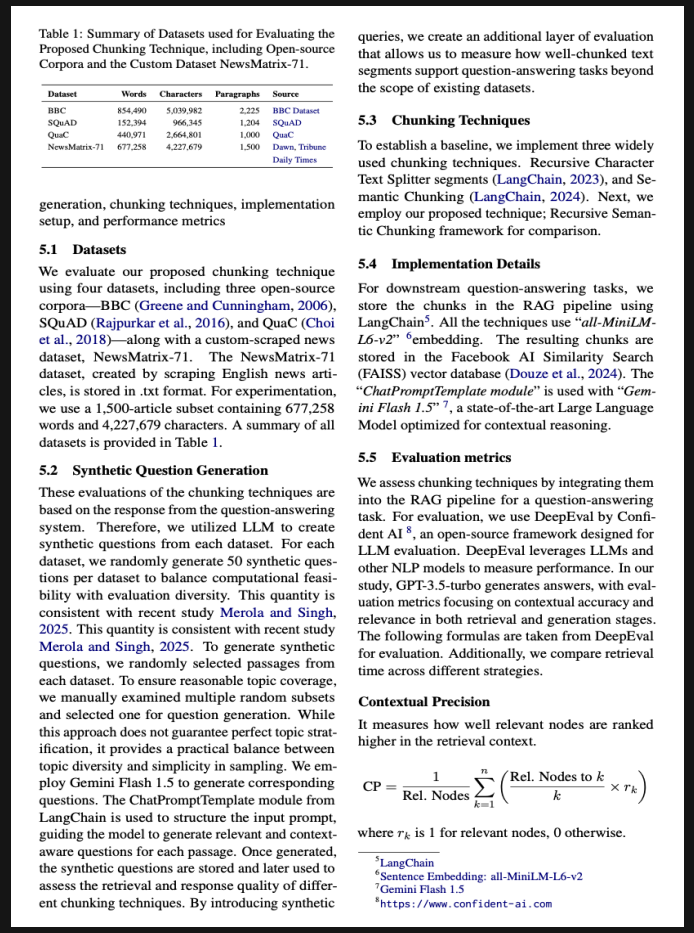


Table 1: Summary of Datasets used for Evaluating the Proposed Chunking Technique, including Open-source Corpora and the Custom Dataset NewsMatrix-71.

Dataset	Words	Characters	Paragraphs	Source
BBC	854,490	5,039,982	2,225	BBC Dataset
SQuAD	152,394	966,345	1,204	SQuAD
QuaC	440,971	2,664,801	1,000	QuaC
NewsMatrix-71	677,258	4,227,679	1,500	Dawn, Tribune Daily Times

generation, chunking techniques, implementation setup, and performance metrics

5.1 Datasets

We evaluate our proposed chunking technique using four datasets, including three open-source corpora—BBC (Greene and Cunningham, 2006), SQuAD (Rajpurkar et al., 2016), and QuaC (Choi et al., 2018)—along with a custom-scraped news dataset, NewsMatrix-71. The NewsMatrix-71 dataset, created by scraping English news articles, is stored in .txt format. For experimentation, we use a 1,500-article subset containing 677,258 words and 4,227,679 characters. A summary of all datasets is provided in Table 1.

5.2 Synthetic Question Generation

These evaluations of the chunking techniques are based on the response from the question-answering system. Therefore, we utilized LLM to create synthetic questions from each dataset. For each dataset, we randomly generate 50 synthetic questions per dataset to balance computational feasibility with evaluation diversity. This quantity is consistent with recent study Merola and Singh, 2025. This quantity is consistent with recent study Merola and Singh, 2025. To generate synthetic questions, we randomly selected passages from each dataset. To ensure reasonable topic coverage, we manually examined multiple random subsets and selected one for question generation. While this approach does not guarantee perfect topic stratification, it provides a practical balance between topic diversity and simplicity in sampling. We employ Gemini Flash 1.5 to generate corresponding questions. The ChatPromptTemplate module from LangChain is used to structure the input prompt, guiding the model to generate relevant and context-aware questions for each passage. Once generated, the synthetic questions are stored and later used to assess the retrieval and response quality of different chunking techniques. By introducing synthetic queries, we create an additional layer of evaluation that allows us to measure how well-chunked text segments support question-answering tasks beyond the scope of existing datasets.

5.3 Chunking Techniques

To establish a baseline, we implement three widely used chunking techniques. Recursive Character Text Splitter segments (LangChain, 2023), and Semantic Chunking (LangChain, 2024). Next, we employ our proposed technique; Recursive Semantic Chunking framework for comparison.

5.4 Implementation Details

For downstream question-answering tasks, we store the chunks in the RAG pipeline using LangChain[5]. All the techniques use "*all-MiniLM-L6-v2*" [6]embedding. The resulting chunks are stored in the Facebook AI Similarity Search (FAISS) vector database (Douze et al., 2024). The "*ChatPromptTemplate module*" is used with "*Gemini Flash 1.5*" [7], a state-of-the-art Large Language Model optimized for contextual reasoning.

5.5 Evaluation metrics

We assess chunking techniques by integrating them into the RAG pipeline for a question-answering task. For evaluation, we use DeepEval by Confident AI [8], an open-source framework designed for LLM evaluation. DeepEval leverages LLMs and other NLP models to measure performance. In our study, GPT-3.5-turbo generates answers, with evaluation metrics focusing on contextual accuracy and relevance in both retrieval and generation stages. The following formulas are taken from DeepEval for evaluation. Additionally, we compare retrieval time across different strategies.

Contextual Precision

It measures how well relevant nodes are ranked higher in the retrieval context.

$$\text{CP} = \frac{1}{\text{Rel. Nodes}} \sum_{k=1}^{n} \left(\frac{\text{Rel. Nodes to } k}{k} \times r_k \right)$$

where r_k is 1 for relevant nodes, 0 otherwise.

[5] LangChain
[6] Sentence Embedding: all-MiniLM-L6-v2
[7] Gemini Flash 1.5
[8] https://www.confident-ai.com

LLM이 잘 해줄 것 같은데?

LLM이 해주지 않을까?

Table 1: Summary of Datasets used for Evaluating the Proposed Chunking Technique, including Open-source Corpora and the Custom Dataset NewsMatrix-71.

Dataset	Words	Characters	Paragraphs	Source
BBC	854,490	5,039,982	2,225	BBC Dataset
SQuAD	152,394	966,345	1,204	SQuAD
QuaC	440,971	2,664,801	1,000	QuaC
NewsMatrix-71	677,258	4,227,679	1,500	Dawn, Tribune Daily Times

generation, chunking techniques, implementation setup, and performance metrics

5.1 Datasets

We evaluate our proposed chunking technique using four datasets, including three open-source corpora—BBC (Greene and Cunningham, 2006), SQuAD (Rajpurkar et al., 2016), and QuaC (Choi et al., 2018)—along with a custom-scraped news dataset, NewsMatrix-71. The NewsMatrix-71 dataset, created by scraping English news articles, is stored in .txt format. For experimentation, we use a 1,500-article subset containing 677,258 words and 4,227,679 characters. A summary of all datasets is provided in Table 1.

5.2 Synthetic Question Generation

These evaluations of the chunking techniques are based on the response from the question-answering system. Therefore, we utilized LLM to create synthetic questions from each dataset. For each dataset, we randomly generate 50 synthetic questions per dataset to balance computational feasibility with evaluation diversity. This quantity is consistent with recent study Merola and Singh, 2025. This quantity is consistent with recent study Merola and Singh, 2025. To generate synthetic questions, we randomly selected passages from each dataset. To ensure reasonable topic coverage, we manually examined multiple random subsets and selected one for question generation. While this approach does not guarantee perfect topic stratification, it provides a practical balance between topic diversity and simplicity in sampling. We employ Gemini Flash 1.5 to generate corresponding questions. The ChatPromptTemplate module from LangChain is used to structure the input prompt, guiding the model to generate relevant and context-aware questions for each passage. Once generated, the synthetic questions are stored and later used to assess the retrieval and response quality of different chunking techniques. By introducing synthetic queries, we create an additional layer of evaluation that allows us to measure how well-chunked text segments support question-answering tasks beyond the scope of existing datasets.

5.3 Chunking Techniques

To establish a baseline, we implement three widely used chunking techniques. Recursive Character Text Splitter segments (LangChain, 2023), and Semantic Chunking (LangChain, 2024). Next, we employ our proposed technique; Recursive Semantic Chunking framework for comparison.

5.4 Implementation Details

For downstream question-answering tasks, we store the chunks in the RAG pipeline using LangChain[5]. All the techniques use "*all-MiniLM-L6-v2*" [6]embedding. The resulting chunks are stored in the Facebook AI Similarity Search (FAISS) vector database (Douze et al., 2024). The "*ChatPromptTemplate module*" is used with "*Gemini Flash 1.5*" [7], a state-of-the-art Large Language Model optimized for contextual reasoning.

5.5 Evaluation metrics

We assess chunking techniques by integrating them into the RAG pipeline for a question-answering task. For evaluation, we use DeepEval by Confident AI [8], an open-source framework designed for LLM evaluation. DeepEval leverages LLMs and other NLP models to measure performance. In our study, GPT-3.5-turbo generates answers, with evaluation metrics focusing on contextual accuracy and relevance in both retrieval and generation stages. The following formulas are taken from DeepEval for evaluation. Additionally, we compare retrieval time across different strategies.

Contextual Precision

It measures how well relevant nodes are ranked higher in the retrieval context.

$$\text{CP} = \frac{1}{\text{Rel. Nodes}} \sum_{k=1}^{n} \left(\frac{\text{Rel. Nodes to } k}{k} \times r_k \right)$$

where r_k is 1 for relevant nodes, 0 otherwise.

[5] LangChain
[6] Sentence Embedding: all-MiniLM-L6-v2
[7] Gemini Flash 1.5
[8] https://www.confident-ai.com

"해줘"

"Chunking해줘"

LLM이 해주지 않을까?

Table 1: Summary of Datasets used for Evaluating the Proposed Chunking Technique, including Open-source Corpora and the Custom Dataset NewsMatrix-71.

Dataset	Words	Characters	Paragraphs	Source
BBC	854,490	5,039,982	2,225	BBC Dataset
SQuAD	152,394	966,345	1,204	SQuAD
QuaC	440,971	2,664,801	1,000	QuaC
NewsMatrix-71	677,258	4,227,679	1,500	Dawn, Tribune Daily Times

generation, chunking techniques, implementation setup, and performance metrics

5.1 Datasets

We evaluate our proposed chunking technique using four datasets, including three open-source corpora—BBC (Greene and Cunningham, 2006), SQuAD (Rajpurkar et al., 2016), and QuaC (Choi et al., 2018)—along with a custom-scraped news dataset, NewsMatrix-71. The NewsMatrix-71 dataset, created by scraping English news articles, is stored in .txt format. For experimentation, we use a 1,500-article subset containing 677,258 words and 4,227,679 characters. A summary of all datasets is provided in Table 1.

5.2 Synthetic Question Generation

These evaluations of the chunking techniques are based on the response from the question-answering system. Therefore, we utilized LLM to create synthetic questions from each dataset. For each dataset, we randomly generate 50 synthetic questions per dataset to balance computational feasibility with evaluation diversity. This quantity is consistent with recent study Merola and Singh, 2025. This quantity is consistent with recent study Merola and Singh, 2025. To generate synthetic questions, we randomly selected passages from each dataset. To ensure reasonable topic coverage, we manually examined multiple random subsets and selected one for question generation. While this approach does not guarantee perfect topic stratification, it provides a practical balance between topic diversity and simplicity in sampling. We employ Gemini Flash 1.5 to generate corresponding questions. The ChatPromptTemplate module from LangChain is used to structure the input prompt, guiding the model to generate relevant and context-aware questions for each passage. Once generated, the synthetic questions are stored and later used to assess the retrieval and response quality of different chunking techniques. By introducing synthetic queries, we create an additional layer of evaluation that allows us to measure how well-chunked text segments support question-answering tasks beyond the scope of existing datasets.

5.3 Chunking Techniques

To establish a baseline, we implement three widely used chunking techniques. Recursive Character Text Splitter segments (LangChain, 2023), and Semantic Chunking (LangChain, 2024). Next, we employ our proposed technique; Recursive Semantic Chunking framework for comparison.

5.4 Implementation Details

For downstream question-answering tasks, we store the chunks in the RAG pipeline using LangChain[5]. All the techniques use "*all-MiniLM-L6-v2*" [6]embedding. The resulting chunks are stored in the Facebook AI Similarity Search (FAISS) vector database (Douze et al., 2024). The "*ChatPromptTemplate module*" is used with "*Gemini Flash 1.5*" [7], a state-of-the-art Large Language Model optimized for contextual reasoning.

5.5 Evaluation metrics

We assess chunking techniques by integrating them into the RAG pipeline for a question-answering task. For evaluation, we use DeepEval by Confident AI [8], an open-source framework designed for LLM evaluation. DeepEval leverages LLMs and other NLP models to measure performance. In our study, GPT-3.5-turbo generates answers, with evaluation metrics focusing on contextual accuracy and relevance in both retrieval and generation stages. The following formulas are taken from DeepEval for evaluation. Additionally, we compare retrieval time across different strategies.

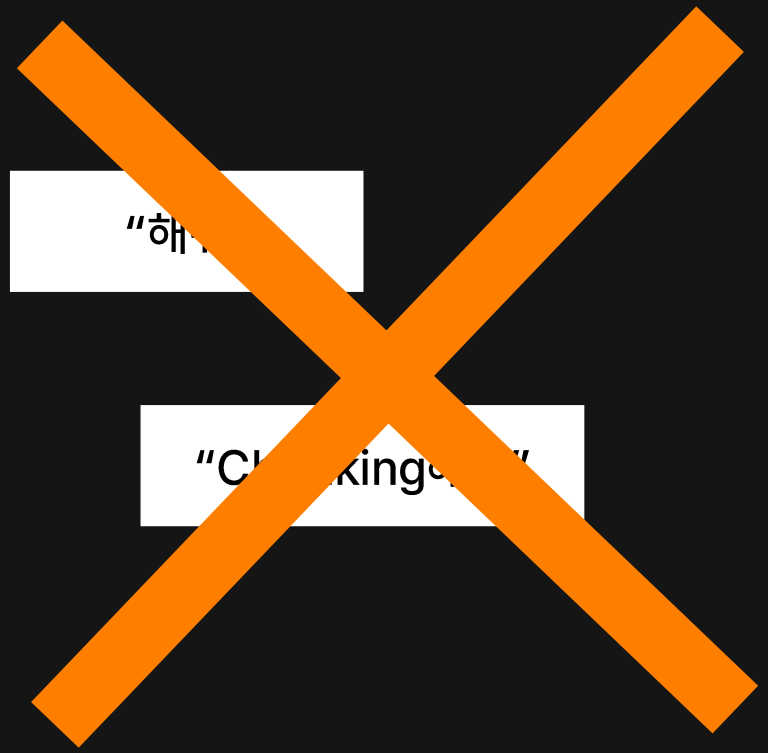
Contextual Precision

It measures how well relevant nodes are ranked higher in the retrieval context.

$$\text{CP} = \frac{1}{\text{Rel. Nodes}} \sum_{k=1}^{n} \left(\frac{\text{Rel. Nodes to } k}{k} \times r_k \right)$$

where r_k is 1 for relevant nodes, 0 otherwise.

[5] LangChain
[6] Sentence Embedding: all-MiniLM-L6-v2
[7] Gemini Flash 1.5
[8] https://www.confident-ai.com



LLM이 해주지 않을까?

안되는 이유가 있음

정보 이론 관점

- Chunking은 전역 최적화 문제 (숲)
- LLM은 Local 확률 모델 (나무)

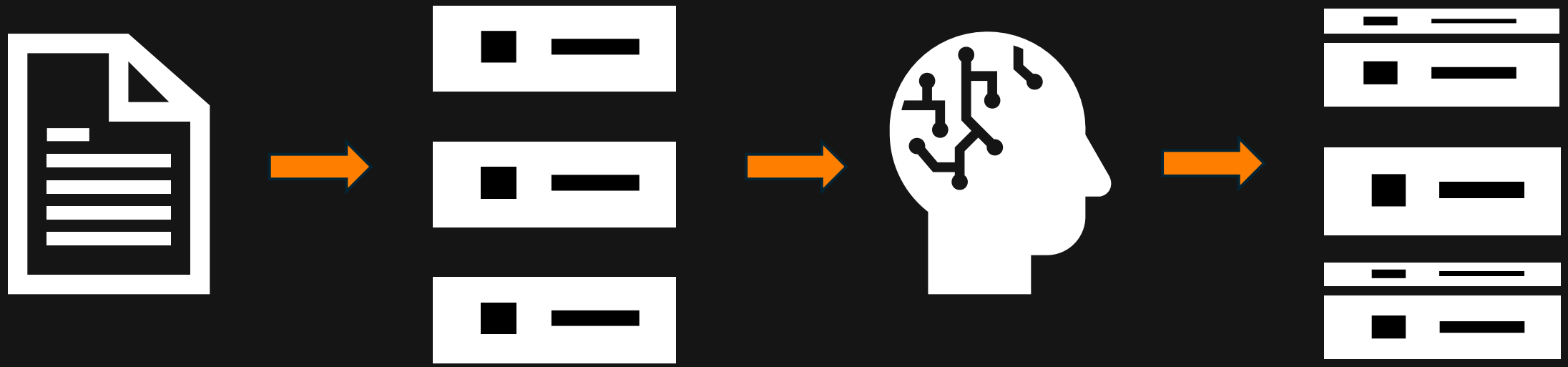
Model 관점

- Transformer는 Next Token Prediction
- 주어진 문맥에서 그럴 듯한 연속을 생성하도록 학습
- 즉, LLM은 각 위치에 대해 독립적인 Label을 생성하게끔 학습 X

- 특히, Context Window가 길어질 수록 Attention이 희석됨

그러면..?

차선책: 특정 단위로 쪼개어서 LLM에 하나씩 넣고 요청하기
그걸 반복하기



LLM의 Local 확률 모델 특성을 더 살릴 수 있게

이 방법도...

한계점 존재

정보 이론 관점

- Chunking은 전역 최적화 문제
- LLM은 Local 확률 모델

실무 적용 관점

- 자체 LLM Model을 가지고 있지 않는 이상 비용 발생

PART 04

구조를 학습시키는 모델링 접근

"알잘딱깔센"

04

문서 넣으면 알아서 구조화 해줄 수 있는 거 없나

Data

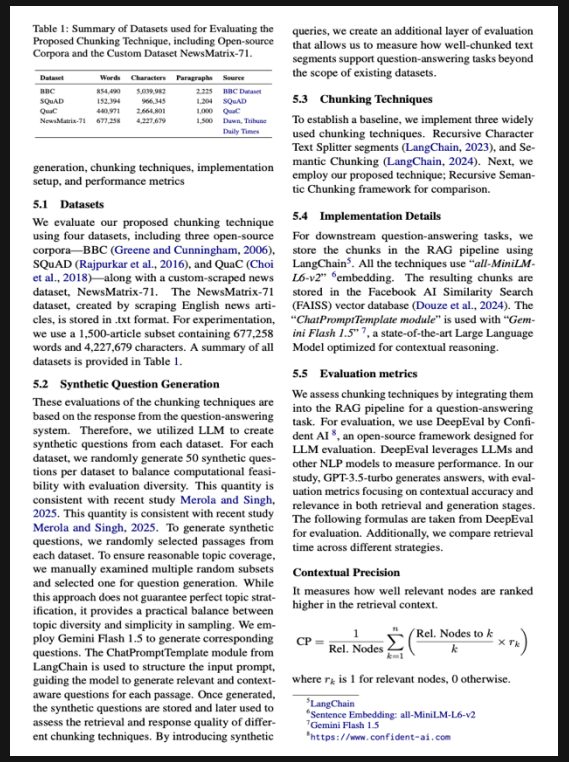


Table 1: Summary of Datasets used for Evaluating the Proposed Chunking Technique, including Open-source Corpora and the Custom Dataset NewsMatrix-71.

Dataset	Words	Characters	Paragraphs	Source
BBC	854,490	5,039,982	2,225	BBC Dataset
SQuAD	152,394	966,345	1,204	SQuAD
QuaC	440,971	2,664,801	1,000	QuaC
NewsMatrix-71	677,258	4,227,679	1,500	Dawn, Tribune Daily Times

generation, chunking techniques, implementation setup, and performance metrics

5.1 Datasets

We evaluate our proposed chunking technique using four datasets, including three open-source corpora—BBC (Greene and Cunningham, 2006), SQuAD (Rajpurkar et al., 2016), and QuaC (Choi et al., 2018)—along with a custom-scraped news dataset, NewsMatrix-71. The NewsMatrix-71 dataset, created by scraping English news articles, is stored in .txt format. For experimentation, we use a 1,500-article subset containing 677,258 words and 4,227,679 characters. A summary of all datasets is provided in Table 1.

5.2 Synthetic Question Generation

These evaluations of the chunking techniques are based on the response from the question-answering system. Therefore, we utilized LLM to create synthetic questions from each dataset. For each dataset, we randomly generate 50 synthetic questions per dataset to balance computational feasibility with evaluation diversity. This quantity is consistent with recent study Merola and Singh, 2025. This quantity is consistent with recent study Merola and Singh, 2025. To generate synthetic questions, we randomly selected passages from each dataset. To ensure reasonable topic coverage, we manually examined multiple random subsets and selected one for question generation. While this approach does not guarantee perfect topic stratification, it provides a practical balance between topic diversity and simplicity in sampling. We employ Gemini Flash 1.5 to generate corresponding questions. The ChatPromptTemplate module from LangChain is used to structure the input prompt, guiding the model to generate relevant and context-aware questions for each passage. Once generated, the synthetic questions are stored and later used to assess the retrieval and response quality of different chunking techniques. By introducing synthetic queries, we create an additional layer of evaluation that allows us to measure how well-chunked text segments support question-answering tasks beyond the scope of existing datasets.

5.3 Chunking Techniques

To establish a baseline, we implement three widely used chunking techniques. Recursive Character Text Splitter segments (LangChain, 2023), and Semantic Chunking (LangChain, 2024). Next, we employ our proposed technique; Recursive Semantic Chunking framework for comparison.

5.4 Implementation Details

For downstream question-answering tasks, we store the chunks in the RAG pipeline using LangChain[5]. All the techniques use "*all-MiniLM-L6-v2*" [6]embedding. The resulting chunks are stored in the Facebook AI Similarity Search (FAISS) vector database (Douze et al., 2024). The "*ChatPromptTemplate module*" is used with "*Gemini Flash 1.5*" [7], a state-of-the-art Large Language Model optimized for contextual reasoning.

5.5 Evaluation metrics

We assess chunking techniques by integrating them into the RAG pipeline for a question-answering task. For evaluation, we use DeepEval by Confident AI [8], an open-source framework designed for LLM evaluation. DeepEval leverages LLMs and other NLP models to measure performance. In our study, GPT-3.5-turbo generates answers, with evaluation metrics focusing on contextual accuracy and relevance in both retrieval and generation stages. The following formulas are taken from DeepEval for evaluation. Additionally, we compare retrieval time across different strategies.

Contextual Precision

It measures how well relevant nodes are ranked higher in the retrieval context.

$$\text{CP} = \frac{1}{\text{Rel. Nodes}} \sum_{k=1}^{n} \left(\frac{\text{Rel. Nodes to } k}{k} \times r_k \right)$$

where r_k is 1 for relevant nodes, 0 otherwise.

[5]LangChain
[6]Sentence Embedding: all-MiniLM-L6-v2
[7]Gemini Flash 1.5
[8]https://www.confident-ai.com

Model

Result

```
Document
├── …
├── 5.1 Datasets
├── 5.2 Synthetic Question Generation
├── 5.3 Chunking Techniques
├── 5.4 Implementation Details
├── 5.5 Evaluation Metrics
└── …
```

Train

데이터

Towards Robust Diabetes-Specialized LLMs: Reflection- and Curriculum-based Instruction Tuning Across Diverse Tasks

Jaesung Hwang, MSc[1] and Donghyeon Park, PhD[2]

[1] Department of Software, Sejong University, Seoul 05006, Republic of Korea.
[2] Department of Artificial Intelligence & Data Science, Sejong University, Seoul 05006, Republic of Korea.

Abstract

Objective: Diabetes is a globally prevalent and complex chronic condition that requires continuous interpretation of glycemic trends, personalized lifestyle interventions, and sustained patient education. While large language models (LLMs) have shown promise in various biomedical applications, existing general-purpose and biomedical LLMs often underperform in diabetes-specific reasoning and instruction-following tasks. To address this limitation, we present a diabetes-specialized LLM fine-tuned on a curated instruction dataset that reflects the nuances of diabetes management.

Methods: Our approach integrates two reflection-based replacement metrics—Instruction-Following Difficulty (IFD) and reversed-IFD (r-IFD)—to evaluate the model's receptiveness to instructions and to filter and substitute suboptimal prompts. In addition, we apply a curriculum instruction tuning strategy that sequences instructions from easier to more challenging based on model sensitivity, mitigating catastrophic forgetting while promoting stable acquisition of domain expertise.

Results: We evaluate the model across multiple diabetes-related task categories, including question answering, natural language inference, information extraction, summarization, generation, and diabetes-friendly dietary recommendation. Our model demonstrates superior performance in accuracy compared to strong biomedical baselines, and generates meal plans that are 0.66% more suitable than those produced by GPT-4, as measured by the Diet Quality Index–International (DQI-I)—an international metric for nutritional quality. Furthermore, simulation-based evaluations using simglucose reveal that our model's meal plans lead to more favorable glycemic outcomes compared to GPT-4.

Conclusion: These results highlight the potential of domain-specific instruction tuning to transform general-purpose LLMs into expert-level assistants for diabetes management.

Key words: Diabetes Management, Reflection-based Replacement, Instruction Following Difficulty, Reversed-Instruction Following Difficulty, Curriculum Instruction Tuning, diabetes-friendly Nutrition, Glycemic Simulation

Introduction

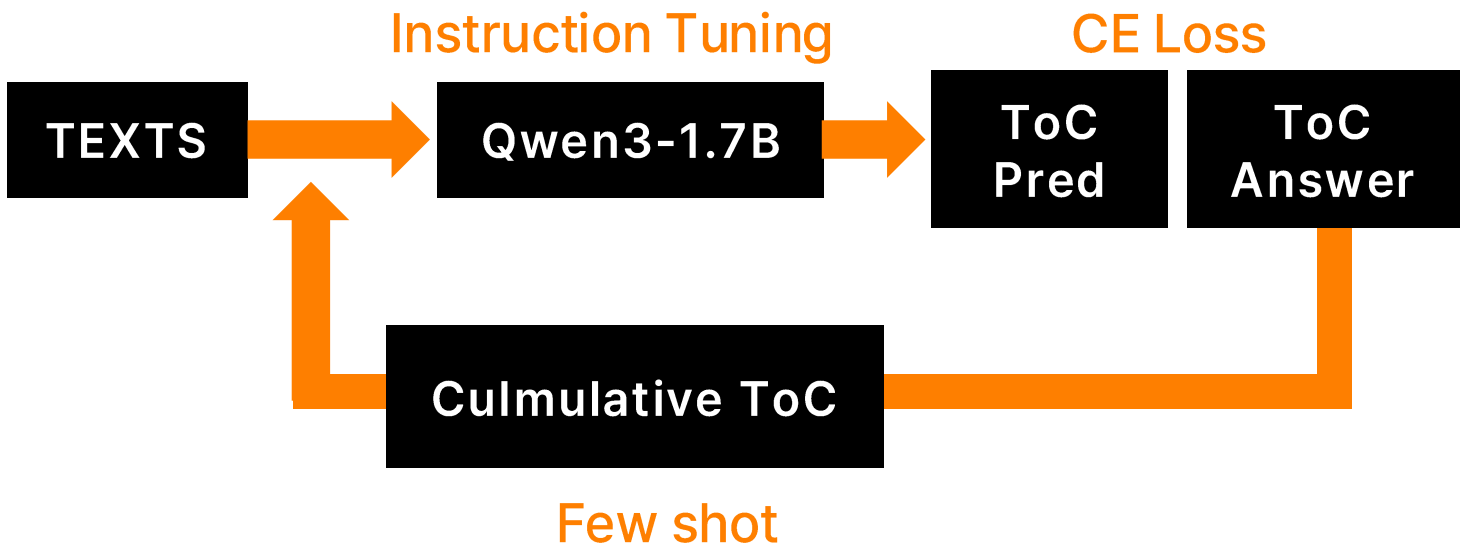
Diabetes is one of the most prevalent and complex chronic diseases globally, currently affecting more than 500 million individuals and contributing significantly to the global burden of non-communicable diseases. According to a comprehensive pooled analysis of 1108 population-representative studies including over 141 million participants, the worldwide prevalence of diabetes has risen continuously from 1990 to 2022, and it is projected to escalate even further in the next decades[25]. Managing diabetes effectively requires not only pharmacological treatment but also ongoing interpretation of blood glucose fluctuations, dietary adaptations, personalized lifestyle interventions, and continuous patient education. As the demand for scalable, patient-centered care rises, the healthcare community increasingly seeks robust AI tools, such as large language models (LLMs), that can support clinical decision-making and self-management guidance.

LLMs have recently gained traction in the biomedical domain due to their capability to generate human-like responses and synthesize complex information. In the context of diabetes management, large language models (LLMs) have been leveraged across a diverse set of tasks tailored to patient needs. For instance, integrated systems combining image-based data with LLM-generated responses have been shown to enhance interpretability in primary management settings[11]. Retrieval-augmented generation (RAG) architectures have been employed to support diabetes education by grounding LLM outputs in trusted reference sources, thereby improving factual reliability[17]. In another example, large language models have been applied to analyze CGM data directly and summarize trends over time for patients and clinicians[1]. But, there is no practical diabetes management LLM that answers questions about diabetes and even provides diet recommendations.

Also, these promising directions still face critical limitations. Biomedical LLMs such as BioBERT[5] and BioGPT[12] have demonstrated high performance on general biomedical QA benchmarks but often underperform on reasoning tasks related to diabetes management, which require clinical nuance, contextual understanding, and personalized judgment. This limitation was further substantiated in a recent prospective study, where even advanced biomedical LLMs failed to achieve satisfactory accuracy and alignment when evaluated on Multiple diabetes-related task. Similarly, general-purpose LLMs like ChatGPT and Bard continue to struggle with endocrinology-related queries, frequently providing vague or partially incorrect answers that lack alignment with clinical guidelines[4]. These limitations underscore the need for models that are not only trained on diabetes-specific data but also optimized for diabetes-specific reasoning and instruction-following in diabetes management, including accurate answers to diabetes-related questions and diabetes-friendly dietary recommendations.

To address this gap, we propose a diabetes-specialized large language model trained on a diabetes-specific instruction dataset designed to reflect the nuances of diabetes management. The dataset includes a diverse set of prompts covering glycemic trend interpretation, insulin dose adjustment, nutritional guidance, symptom analysis, and patient counseling. Importantly, the model is trained not merely to recall factual information but to engage in higher-order reasoning that emulates the decision-making patterns of healthcare professionals. To support this goal, we go beyond merely

아키텍쳐



Train

한계점...

정보 이론 관점

- Chunking은 **전역 최적화 문제**
- LLM은 **Local 확률 모델**

Train

한계점 **개선!!**

정보 이론 관점

- Chunking은 전역 최적화 문제
- LLM은 Local 확률 모델 **+ 전역 정보**

Train

데이터


```
{
    "tree": "",
    "input": "Towards Robust Diabetes-Specialized LLMs: Reflection- and\nCurricu
    "output": "+\n*\n"
},
{
    "tree": "+ Towards Robust Diabetes-Specialized LLMs...\n",
    "input": "Jaesung Hwang, MSc[1] and Donghyeon Park, PhD[2]\nAbstract\nObjective:
    "output": "*\n+\n*\n"
},
{
    "tree": "+ Towards Robust Diabetes-Specialized LLMs...\n+ Abstract\n",
    "input": "Objective: Diabetes is a globally prevalent and complex\nchronic c
    "output": "*\n=\n=\n"
},
```

Inference

데이터

Towards Robust Diabetes-Specialized LLMs: Reflection- and Curriculum-based Instruction Tuning Across Diverse Tasks

Jaesung Hwang, MSc[1] and Donghyeon Park, PhD[2]

[1]Department of Software, Sejong University, Seoul 05006, Republic of Korea.
[2]Department of Artificial Intelligence & Data Science, Sejong University, Seoul 05006, Republic of Korea.

Abstract

Objective: Diabetes is a globally prevalent and complex chronic condition that requires continuous interpretation of glycemic trends, personalized lifestyle interventions, and sustained patient education. While large language models (LLMs) have shown promise in various biomedical applications, existing general-purpose and biomedical LLMs often underperform in diabetes-specific reasoning and instruction-following tasks. To address this limitation, we present a diabetes-specialized LLM fine-tuned on a curated instruction dataset that reflects the nuances of diabetes management.

Methods: Our approach integrates two reflection-based replacement metrics—Instruction-Following Difficulty (IFD) and reversed-IFD (r-IFD)—to evaluate the model's receptiveness to instructions and to filter and substitute suboptimal prompts. In addition, we apply a curriculum instruction tuning strategy that sequences instructions from easier to more challenging based on model sensitivity, mitigating catastrophic forgetting while promoting stable acquisition of domain expertise.

Results: We evaluate the model across multiple diabetes-related task categories, including question answering, natural language inference, information extraction, summarization, generation, and diabetes-friendly dietary recommendation. Our model demonstrates superior performance in accuracy compared to strong biomedical baselines, and generates meal plans that are 0.66% more suitable than those produced by GPT-4, as measured by the Diet Quality Index–International (DQI-I)—an international metric for nutritional quality. Furthermore, simulation-based evaluations using simglucose reveal that our model's meal plans lead to more favorable glycemic outcomes compared to GPT-4.

Conclusion: These results highlight the potential of domain-specific instruction tuning to transform general-purpose LLMs into expert-level assistants for diabetes management.

Key words: Diabetes Management, Reflection-based Replacement, Instruction Following Difficulty, Reversed-Instruction Following Difficulty, Curriculum Instruction Tuning, diabetes-friendly Nutrition, Glycemic Simulation

Introduction

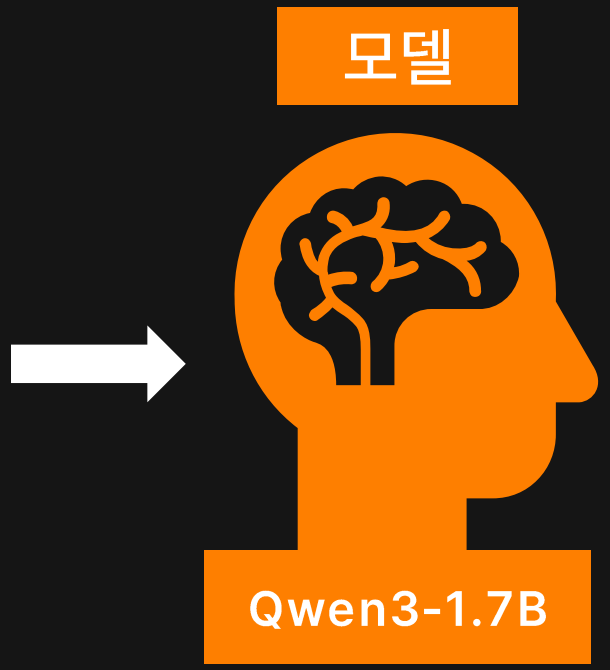
Diabetes is one of the most prevalent and complex chronic diseases globally, currently affecting more than 500 million individuals and contributing significantly to the global burden of non-communicable diseases. According to a comprehensive pooled analysis of 1108 population-representative studies including over 141 million participants, the worldwide prevalence of diabetes has risen continuously from 1990 to 2022, and it is projected to escalate even further in the next decades[20]. Managing diabetes effectively requires not only pharmacological treatment but also ongoing interpretation of blood glucose fluctuations, dietary adaptations, personalized lifestyle interventions, and continuous patient education. As the demand for scalable, patient-centered care rises, the healthcare community increasingly seeks robust AI tools, such as large language models (LLMs), that can support clinical decision-making and self-management guidance.

LLMs have recently gained traction in the biomedical domain due to their capability to generate human-like responses and synthesize complex information. In the context of diabetes management, large language models (LLMs) have been leveraged across a diverse set of tasks tailored to patient needs. For instance, integrated systems combining image-based data with LLM-generated responses have been shown to enhance interpretability in primary management settings[11]. Retrieval-augmented generation (RAG) architectures have been employed to support diabetes education by grounding LLM outputs in trusted reference sources, thereby improving factual reliability[17]. In another example, large language models have been applied to analyze CGM data directly and summarize trends over time for patients and clinicians[4]. But, there is no practical diabetes management LLM that answers questions about diabetes and even provides diet recommendations.

Also, these promising directions still face critical limitations. Biomedical LLMs such as BioBERT[5] and BioGPT[12] have demonstrated high performance on general biomedical QA benchmarks but often underperform on reasoning tasks related to diabetes management, which require clinical nuance, contextual understanding, and personalized judgment. This limitation was further substantiated in a recent prospective study, where even advanced biomedical LLMs failed to achieve satisfactory accuracy and alignment when evaluated on Multiple diabetes-related task. Similarly, general-purpose LLMs like ChatGPT and Bard continue to struggle with endocrinology-related queries, frequently providing vague or partially incorrect answers that lack alignment with clinical guidelines[6]. These limitations underscore the need for models that are not only trained on diabetes-specific data but also optimized for diabetes-specific reasoning and instruction-following in diabetes management, including accurate answers to diabetes-related questions and diabetes-friendly dietary recommendations.

To address this gap, we propose a diabetes-specialized large language model trained on a diabetes-specific instruction dataset designed to reflect the nuances of diabetes management. The dataset includes a diverse set of prompts covering glycemic trend interpretation, insulin dose adjustment, nutritional guidance, symptom analysis, and patient counseling. Importantly, the model is trained not merely to recall factual information but to engage in higher-order reasoning that emulates the decision-making patterns of healthcare professionals. To support this goal, we go beyond merely

모델



결과

```
root ROOT
 └── [TITLE]
      ├── 1. INTRODUCTION
      ├── 2. TASK SETUP
      │    ├── 2.1. Description
      │    ├── 2.2. Timeline
      │    ├── 2.3. Development and evaluation dataset
      │    └── 2.4. Baseline system
      ├── 3. BASELINE SYSTEM RESULTS
      ├── 4. CON CLUS IONS
      ├── 5. NOTE
      ├── 6. ACKNOWLEDGMENT
      └── 7. REFERENCES
```

현재 진행 상황

Answer

```
root ROOT
    └─ [TITLE]
        ├─ 1. INTRODUCTION
        ├─ 2. TASK SETUP
            ├─ 2.1. Description
            ├─ 2.2. Timeline
            ├─ 2.3. Development and evalu ation dataset
            └─ 2.4. Baseline system
        ├─ 3. BASELINE SYSTEM RESULTS
        ├─ 4. CON CLUS IONS
        ├─ 5. NOTE
        ├─ 6. ACKNOWLEDGMENT
        └─ 7. REFERENCES
```

Pred

```
root ROOT
    └─ [TITLE] DCASE 2018 CHALLENGE - TASK 5: MONITORING OF DOME...
        ├─ 1. INTRODUCTION
        ├─ 2. TASK SETUP
        ├─ 2.1. Description
            ├─ 2.2. Timeline
            ├─ 2.3. Development and evalu ation dataset
            └─ 2.4. Baseline system
        ├─ 3. BASELINE SYSTEM RESULTS
        ├─ 4. CON CLUS IONS
        ├─ 5. NOTE
        ├─ 6. ACKNOWLEDGMENT
        └─ 7. REFERENCES
```

Precision: 79.773
Recall: 89.842
F1: 84.509
Tree Edit Distance-based Similarity = 70.517

LLM이 데이터를 더 잘 이해하려면 어떻게 해야할지?

PART 05

Lazy Chunking

"근데... 모든 걸 다 미리 Chunking 해놔야할까?"

05

Lazy Chunking

모든 문서에 대해서 Chunking을 해야할까?

Lazy Chunking

어떤 질문을 더 많이 하지?

Document-level 질문

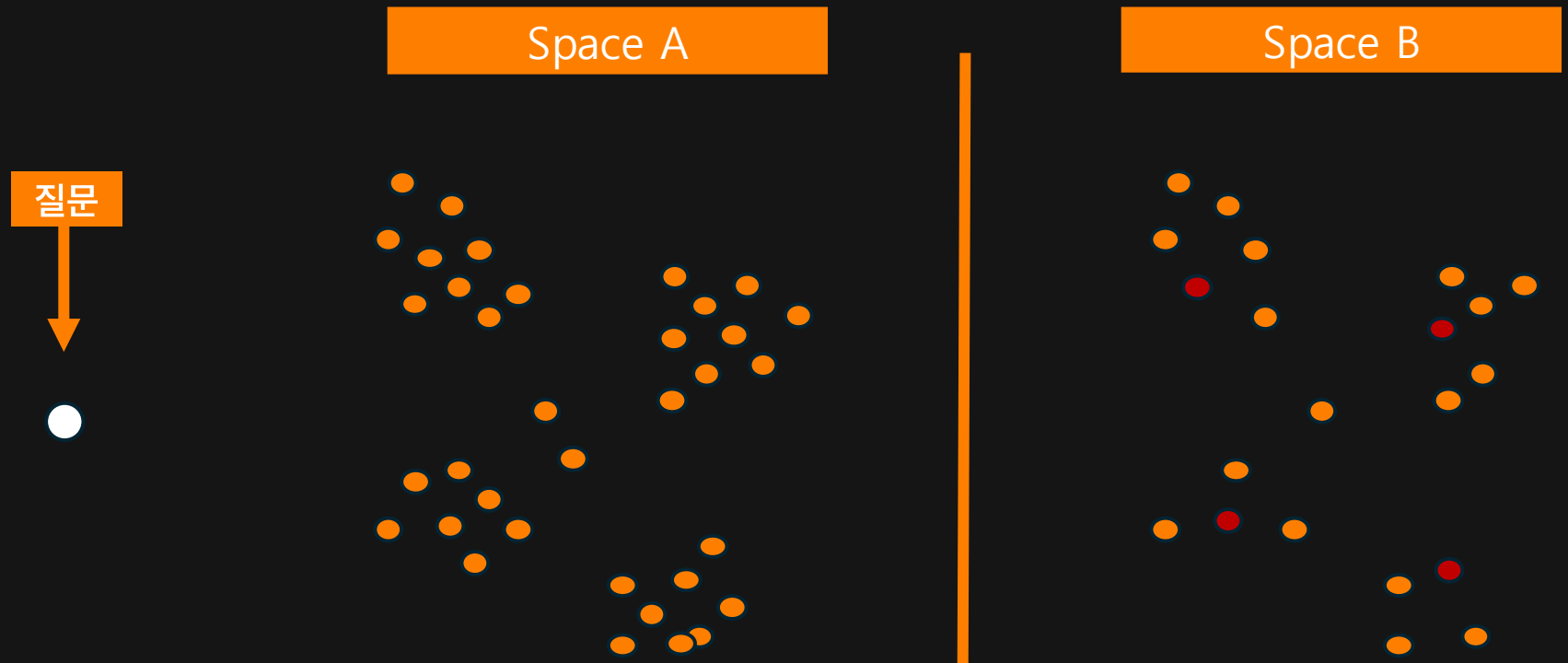
- 휴가 규정은 어느 문서에 정의되어 있나요?
- 외주 계약 관련 기준은 어떤 문서에 있나요?
- 보안 정책과 보안 운영 매뉴얼의 차이는 뭐야?
- 인사 관련 규정의 큰 흐름을 설명해줘.

Chunk-level 질문

- 연차는 최대 몇 일까지 이월할 수 있지?
- 야근 수당은 어떤 경우에 지급할 수 있어?
- 정규직과 계약직의 휴가 규정 차이는?

Lazy Chunking

정보 손실 없는 벡터 공간 슬림화

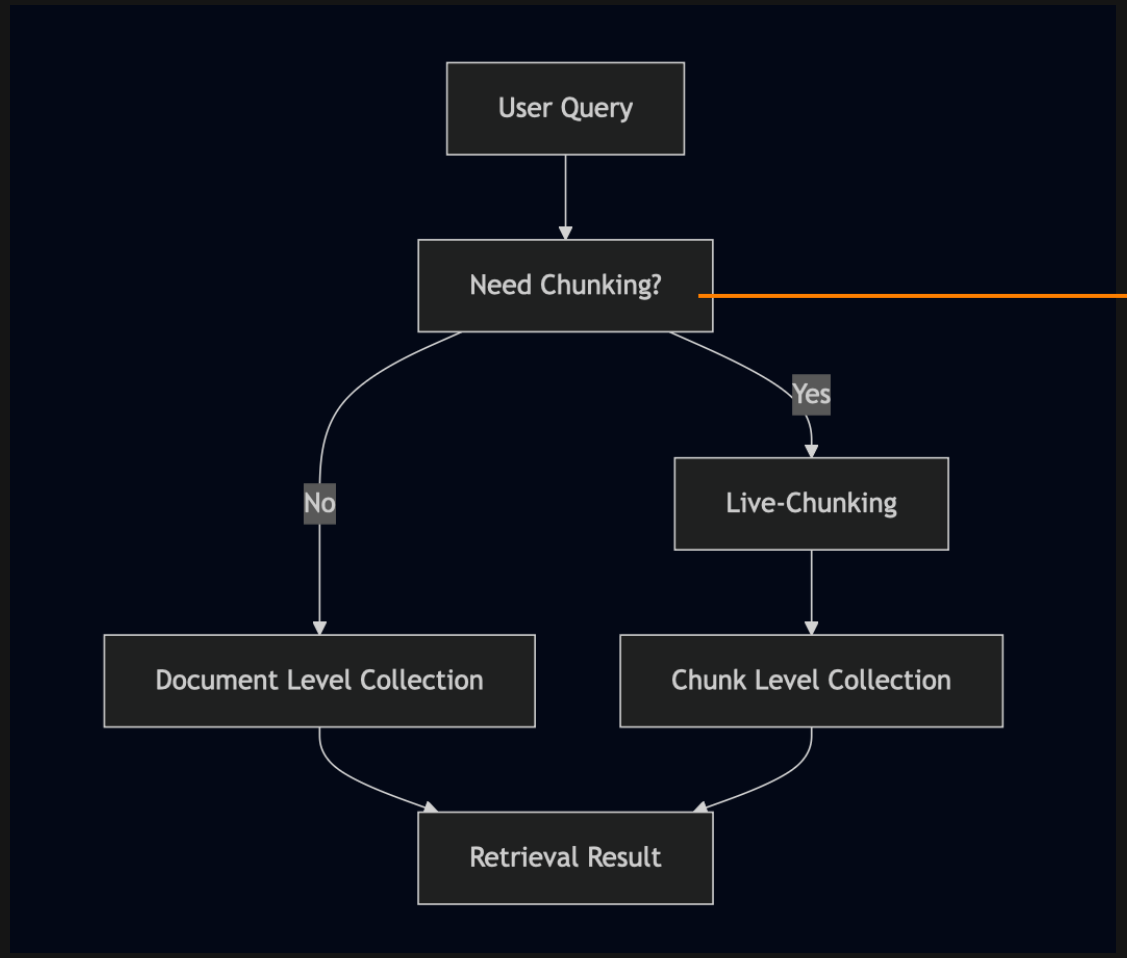


모든 문서가 Chunking 된 경우

꼭 필요한 문서만 Chunking 된 경우
(나머지는 문서 Embedding)

Lazy Chunking

Architecture



- 문서 단위 검색을 우선적으로 적용
- 질문 시점에 해당하는 Document Embedding의 토큰 수가 임계값을 넘는다면, Chunk로 쪼개어서 Chunk 단위 검색 수행

Lazy Chunking

Need Chunking?

기준

현재 - 토큰 수

이유

Lazy Chunking

장단점

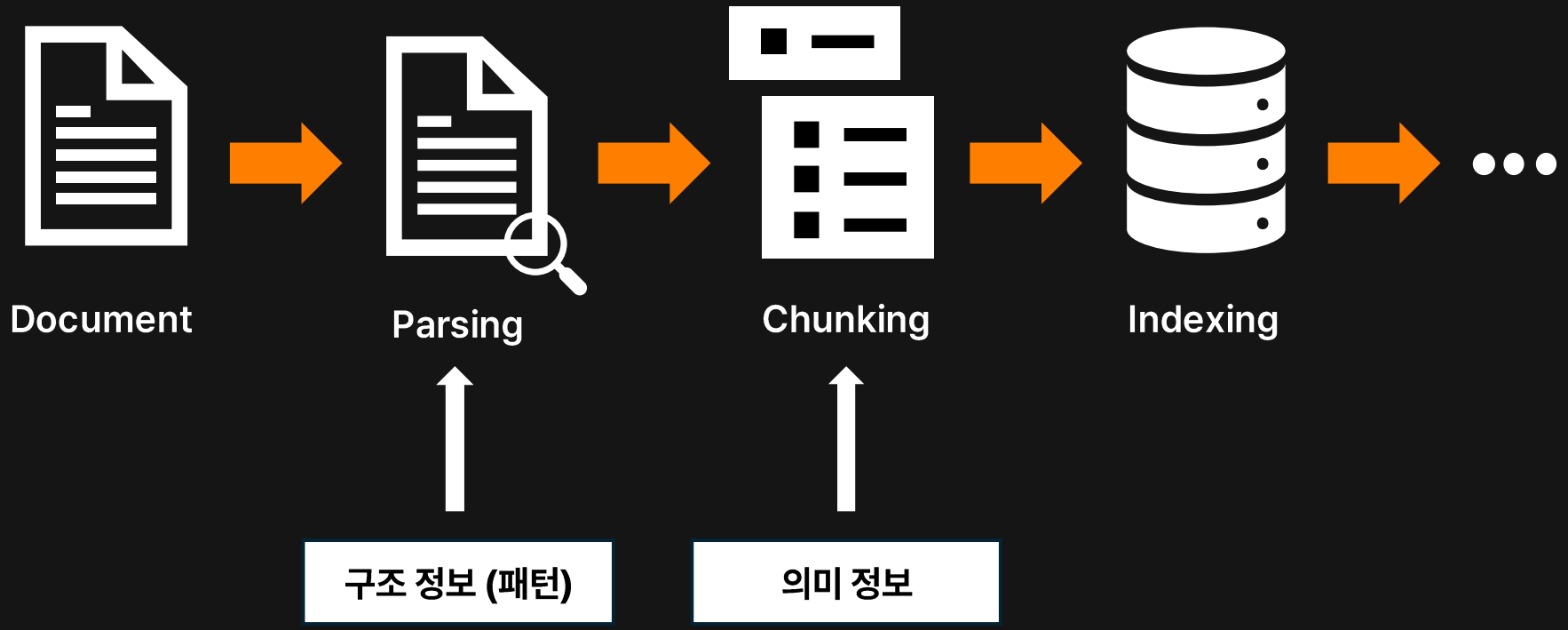
장점

- 메모리 효율적
- 검색 공간의 최적화 가능
    - 보다 높은 정확도 가능
- 원본 맥락 유지 가능
    - 전체 맥락을 파악하기 용이

단점

- Latency 증가할 수 있음
- Embedding Model의 부하

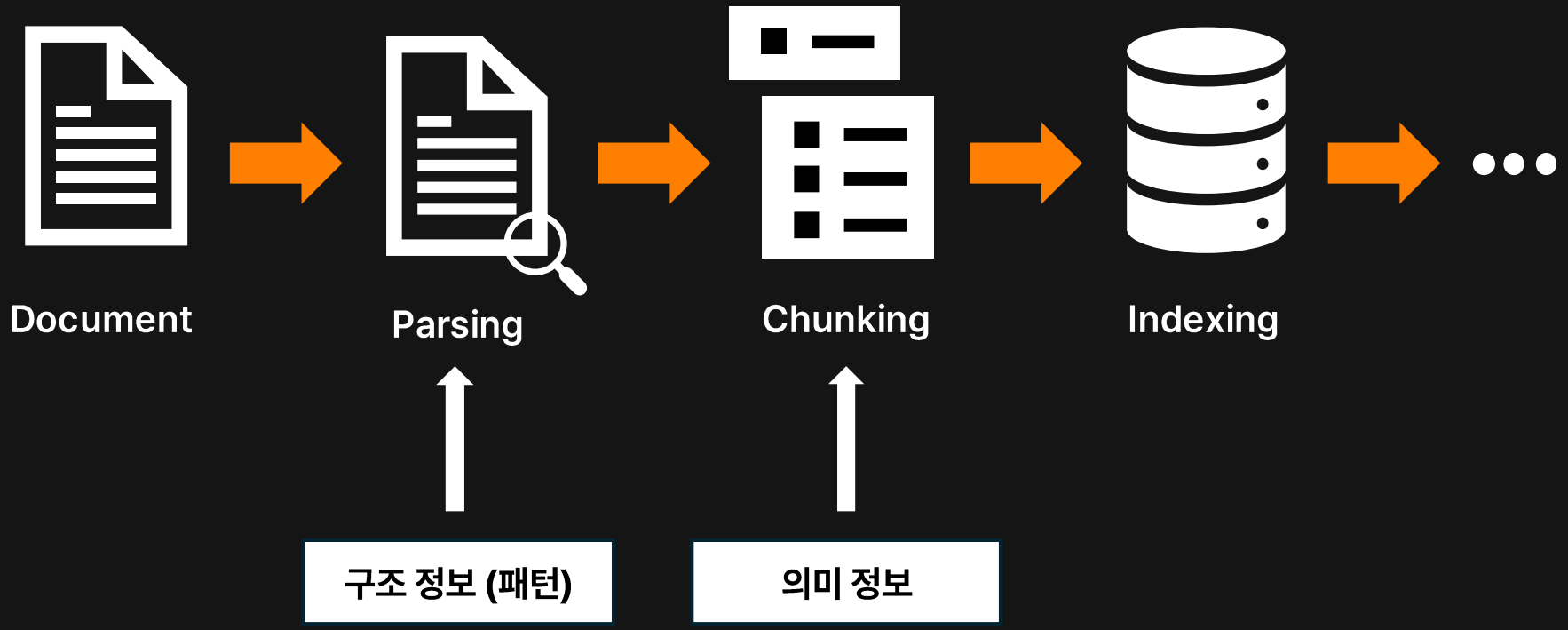
정리



개선점

- Chunking이 중요한데, 앞단의 구조 정보를 잘 활용하지 못함

정리



개선점

- Chunking이 중요한데, 앞단의 구조 정보를 잘 활용하지 못함
- 개선하기 위해서 **Structuring을 도입 중**

예시

데이터


```
{
    "tree": "",
    "input": "Towards Robust Diabetes-Specialized LLMs: Reflection- and\nCurricu
    "output": "+\n*\n"
},
{
    "tree": "+ Towards Robust Diabetes-Specialized LLMs...\n",
    "input": "Jaesung Hwang, MSc¹ and Donghyeon Park, PhD²\nAbstract\nObjective:
    "output": "*\n+\n*\n"
},
{
    "tree": "+ Towards Robust Diabetes-Specialized LLMs...\n+ Abstract\n",
    "input": "Objective: Diabetes is a globally prevalent and complex\nchronic c
    "output": "*\n=\n=\n"
},
```


Towards Robust Diabetes-Specialized LLMs: Reflection- and Curriculum-based Instruction Tuning Across Diverse Tasks

Jaesung Hwang, MSc[1] and Donghyeon Park, PhD[2]

[1]Department of Software, Sejong University, Seoul 05006, Republic of Korea.
[2]Department of Artificial Intelligence & Data Science, Sejong University, Seoul 05006, Republic of Korea.

Abstract

Objective: Diabetes is a globally prevalent and complex chronic condition that requires continuous interpretation of glycemic trends, personalized lifestyle interventions, and sustained patient education. While large language models (LLMs) have shown promise in various biomedical applications, existing general-purpose and biomedical LLMs often underperform in diabetes-specific reasoning and instruction-following tasks. To address this limitation, we present a diabetes-specialized LLM fine-tuned on a curated instruction dataset that reflects the nuances of diabetes management.

Methods: Our approach integrates two reflection-based replacement metrics—Instruction-Following Difficulty (IFD) and reversed-IFD (r-IFD)—to evaluate the model's receptiveness to instructions and to filter and substitute suboptimal prompts. In addition, we apply a curriculum instruction tuning strategy that sequences instructions from easier to more challenging based on model sensitivity, mitigating catastrophic forgetting while promoting stable acquisition of domain expertise.

Results: We evaluate the model across multiple diabetes-related task categories, including question answering, natural language inference, information extraction, summarization, generation, and diabetes-friendly dietary recommendation. Our model demonstrates superior performance in accuracy compared to strong biomedical baselines, and generates meal plans that are 0.66% more suitable than those produced by GPT-4, as measured by the Diet Quality Index–International (DQI-I)—an international metric for nutritional quality. Furthermore, simulation-based evaluations using simglucose reveal that our model's meal plans lead to more favorable glycemic outcomes compared to GPT-4.

Conclusion: These results highlight the potential of domain-specific instruction tuning to transform general-purpose LLMs into expert-level assistants for diabetes management.

Key words: Diabetes Management, Reflection-based Replacement, Instruction Following Difficulty, Reversed-Instruction Following Difficulty, Curriculum Instruction Tuning, diabetes-friendly Nutrition, Glycemic Simulation

Introduction

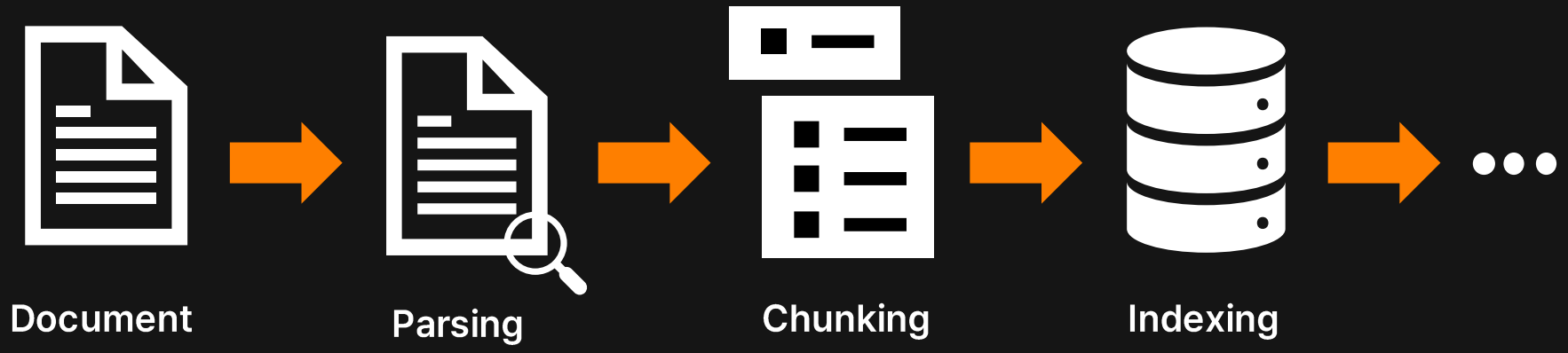
Diabetes is one of the most prevalent and complex chronic diseases globally, currently affecting more than 500 million individuals and contributing significantly to the global burden of non-communicable diseases. According to a comprehensive pooled analysis of 1108 population-representative studies including over 141 million participants, the worldwide prevalence of diabetes has risen continuously from 1990 to 2022, and it is projected to escalate even further in the next decades[25]. Managing diabetes effectively requires not only pharmacological treatment but also ongoing interpretation of blood glucose fluctuations, dietary adaptations, personalized lifestyle interventions, and continuous patient education. As the demand for scalable, patient-centered care rises, the healthcare community increasingly seeks robust AI tools, such as large language models (LLMs), that can support clinical decision-making and self-management guidance.

LLMs have recently gained traction in the biomedical domain due to their capability to generate human-like responses and synthesize complex information. In the context of diabetes management, large language models (LLMs) have been leveraged across a diverse set of tasks tailored to patient needs. For instance, integrated systems combining image-based data with LLM-generated responses have been shown to enhance interpretability in primary management settings[11]. Retrieval-augmented generation (RAG) architectures have been employed to support diabetes education by grounding LLM outputs in trusted reference sources, thereby improving factual reliability[17]. In another example, large language models have been applied to analyze CGM data directly and summarize trends over time for patients and clinicians[1]. But, there is no practical diabetes management LLM that answers questions about diabetes and even provides diet recommendations.

Also, these promising directions still face critical limitations. Biomedical LLMs such as BioBERT[5] and BioGPT[12] have demonstrated high performance on general biomedical QA benchmarks but often underperform on reasoning tasks related to diabetes management, which require clinical nuance, contextual understanding, and personalized judgment. This limitation was further substantiated in a recent prospective study, where even advanced biomedical LLMs failed to achieve satisfactory accuracy and alignment when evaluated on Multiple diabetes-related task. Similarly, general-purpose LLMs like ChatGPT and Bard continue to struggle with endocrinology-related queries, frequently providing vague or partially incorrect answers that lack alignment with clinical guidelines[4]. These limitations underscore the need for models that are not only trained on diabetes-specific data but also optimized for diabetes-specific reasoning and instruction-following in diabetes management, including accurate answers to diabetes-related questions and diabetes-friendly dietary recommendations.

To address this gap, we propose a diabetes-specialized large language model trained on a diabetes-specific instruction dataset designed to reflect the nuances of diabetes management. The dataset includes a diverse set of prompts covering glycemic trend interpretation, insulin dose adjustment, nutritional guidance, symptom analysis, and patient counseling. Importantly, the model is trained not merely to recall factual information but to engage in higher-order reasoning that emulates the decision-making patterns of healthcare professionals. To support this goal, we go beyond merely

BrainCrew Data Team (일부) 방향성



방향성

- Document Chunking은 RAG에서 중요한 역할
  - Document의 시각 정보 또한 의미에 큰 영향을 미침
- 문서에 대한 도메인 전문가는 최적의 Chunking 기법을 찾을 수 있음
  - 구성이 바뀌면 처음부터 다시 설계해야 함
- AI comprehension
  - 이해를 잘 하려면 문서의 구조 정보와 의미 정보를 동시에 다루는 모델이 필요
  - 이게 일반화의 시작점이 될 것

결론

1. RAG의 AI Comprehension을 위해서는 Chunking 기법이 중요하다.

2. Chunking을 위해서는 문장 구조와 의미 구조를 모두 사용해야 한다.

3. BrainCrew Data Team은 Chunking의 일반화를 지향하고 있다.

4. 그 중 일부인 Structuring에 대한 소개를 진행했다.

채용 공고

채용 공고

하는 일

- RAG/AGENT를 위한 Data Processing
  - Document
  - Streaming Data
  - OCR
- Knowledge 내재화
- Data Pipelines
- Agent 개발

도구

- Engineering, Modeling, LLM, ...

조건

- 기술에 매몰되기 보다 문제 정의를 잘 하시는 분
- 기술적 의사결정이 가능하신 분
- 기술에 대한 논의하는 것을 즐기시는 분

Thank You