AX를 위한 Document Parsing

Document Parsing을 더 잘하는 방법

Peter

CONTENTS

PART 01

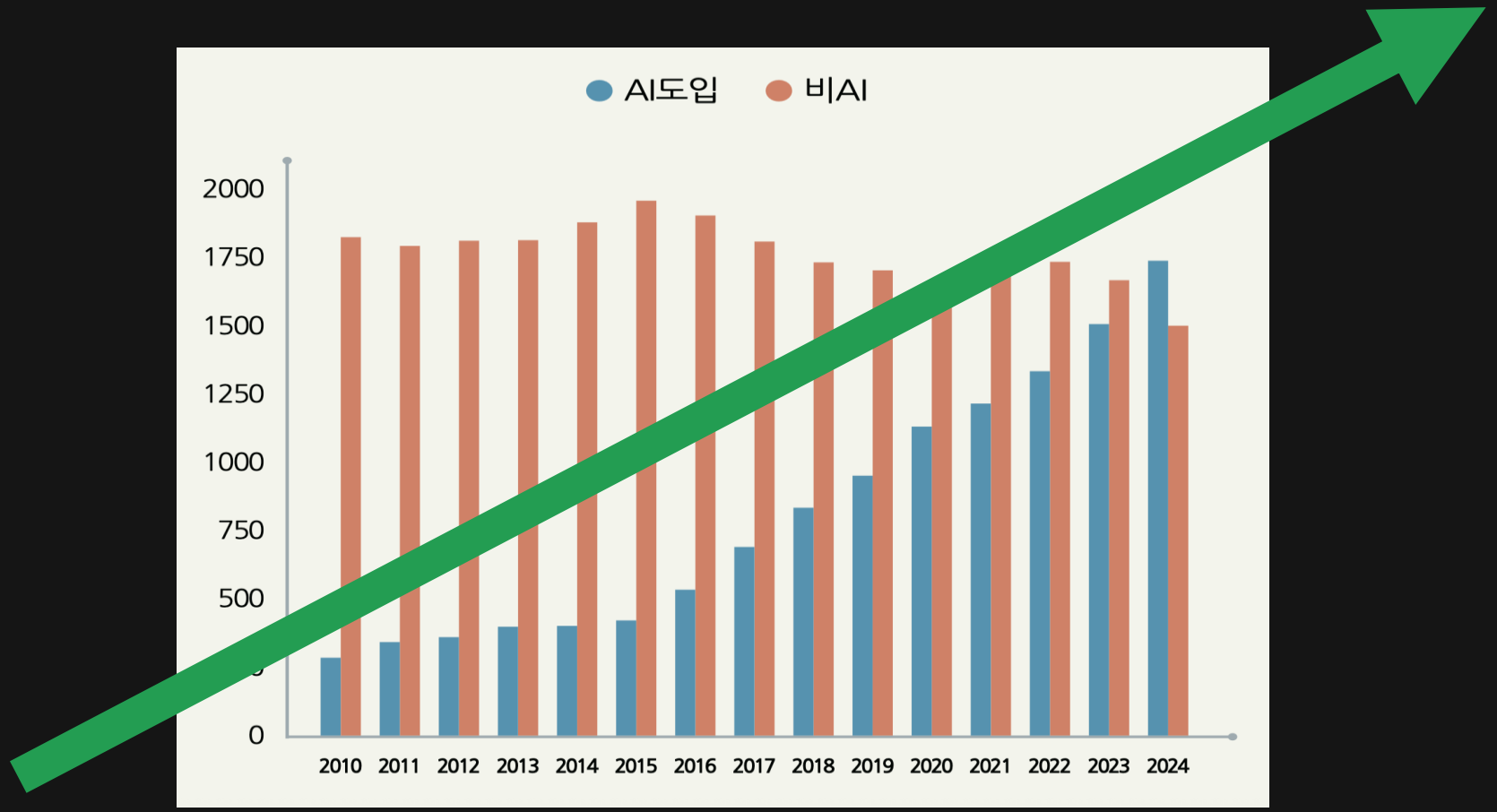
기업들의 AX 도입

"문제는 문서를 다루는 방식이다."

01

기업 AX 도입 트렌드와 현실

기업에서는 AX를 활발히 적용 중



기업 AX 도입 트렌드와 현실

기업에서는 AX를 활발히 적용 중

하지만!

기업 AX 도입 트렌드와 현실

[기고] 산업 AI 도입, 80%가 실패하는 진짜 이유

발행일 : 2025-12-29 16:00 지면 : 2025-12-30 26면

2024년 국내 제조업 인공지능(AI) 도입률이 35%를 넘어섰다. 그러나 랜드연구소(RAND Corporation) 연구에 따르면 AI 프로젝트 80%가 프로덕션 단계에 도달하지 못하고 있다. 일반 정보기술(IT) 프로젝트 실패율의 두 배다.

전시용 데모에서 빛났던 프로젝트가 현장에 투입되는 순간 좌초한다. 15년간 AI 현장에서 목격한 실패들은 한 가지 공통점을 보였다. 문제는 알고리즘이 아니었다. 흩어진 데이터, 불신하는 현장, 그리고 20년 된 시스템과의 단절이었다.

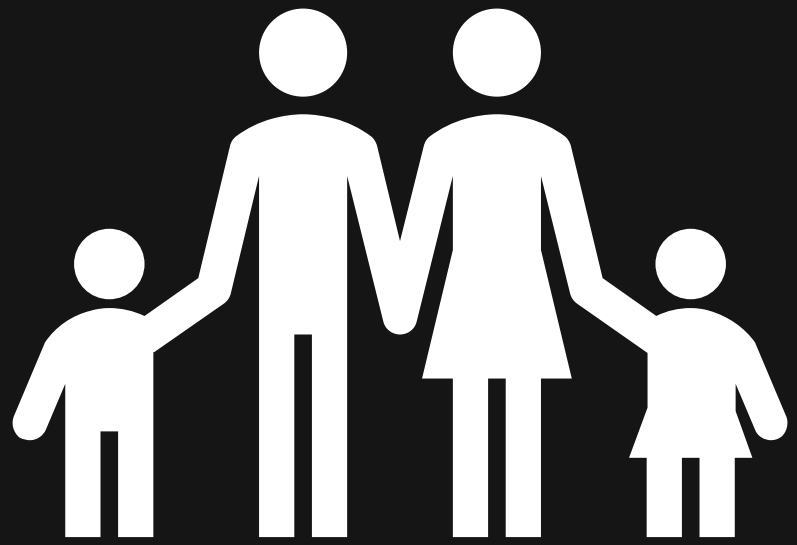
가장 흔한 실패는 데이터 준비의 과소평가다. 최첨단 레이싱카를 사고 연료 품질은 점검하지 않는 격이다. 한 반도체 기업은 수억원짜리 딥러닝 모델을 도입했지만, 20년 된 제조실행시스템(MES)의 데이터를 꺼내는 데만 8개월을 허비했다. 데이터는 시스템마다 흩어진 '사일로' 상태였고, 형식조차 제각각이었다. 더 큰 문제는 데이터의 '의미'였다. '온도 35도'라는 숫자 하나를 보자. 설비 센서 값인지, 제품 온도인지, 외부 기온인지를 사람은 맥락으로 안다. AI는 모른다. 라벨 없는 숫자는 AI에게 외국어와 같다. 60%의 AI 리더가 레거시 통합을 최대 장애물로 꼽는 이유다.

기업 AX 도입 트렌드와 현실

AX를 "잘" 하려면?

신입사원

신입사원이 들어왔다면?



신입사원

신입사원

회사 매뉴얼

신입사원

신입사원이 들어왔다면?

1. 자리에 앉힌다.

2. 매뉴얼을 보라고 한다.

3. 매뉴얼에 대한 질의를 한다.

-3. 신입이 온다고 한다.

-2. 매뉴얼 자료를 검토한다.

-1. 매뉴얼을 이해하기 쉽게 재구성한다.

0. 신입사원을 기다린다.

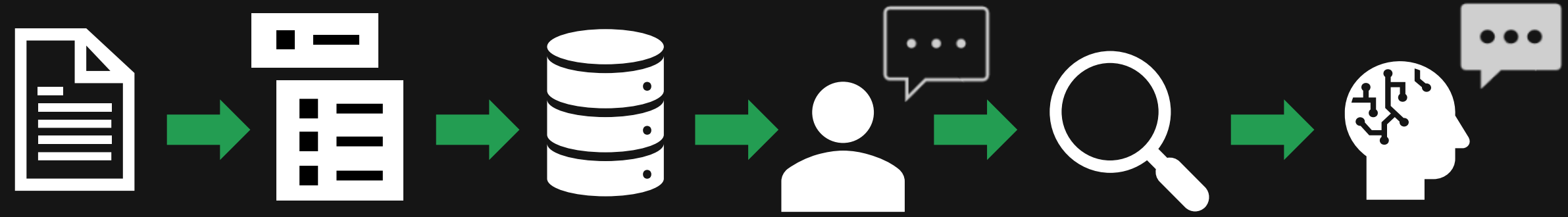
1. 자리에 앉힌다.

2. 매뉴얼을 보라고 한다.

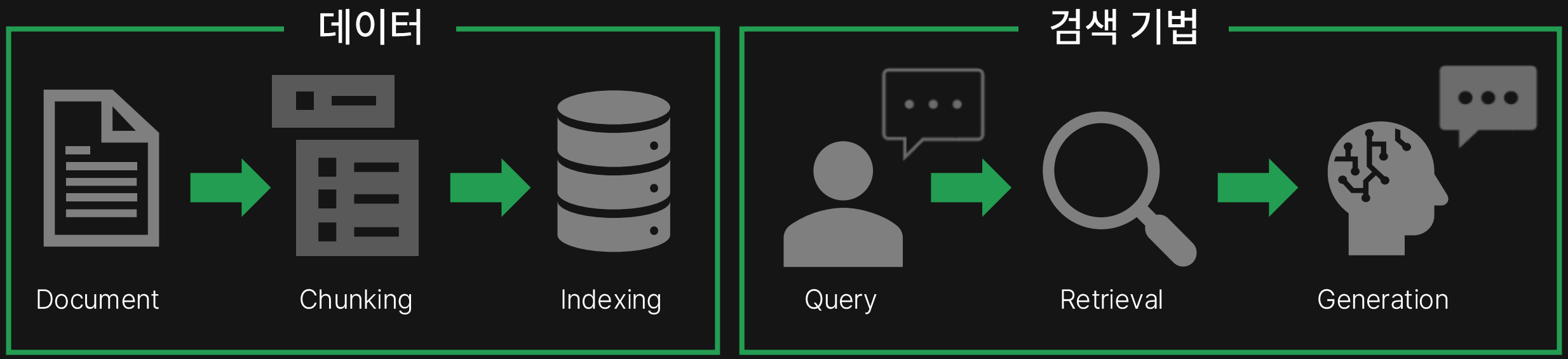
3. 매뉴얼에 대한 질의를 한다.

RAG/AGENT 도입의 기대와 한계

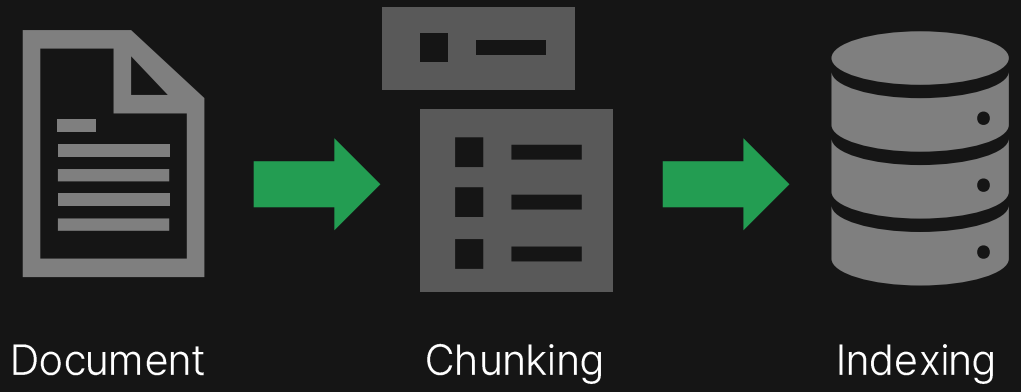
일단, 보편적인 RAG FLOW는 뭐지?



RAG/AGENT 도입의 기대와 한계



데이터가 병목이 되는 이유 1



주로 데이터보다는 검색 기법을 최적화

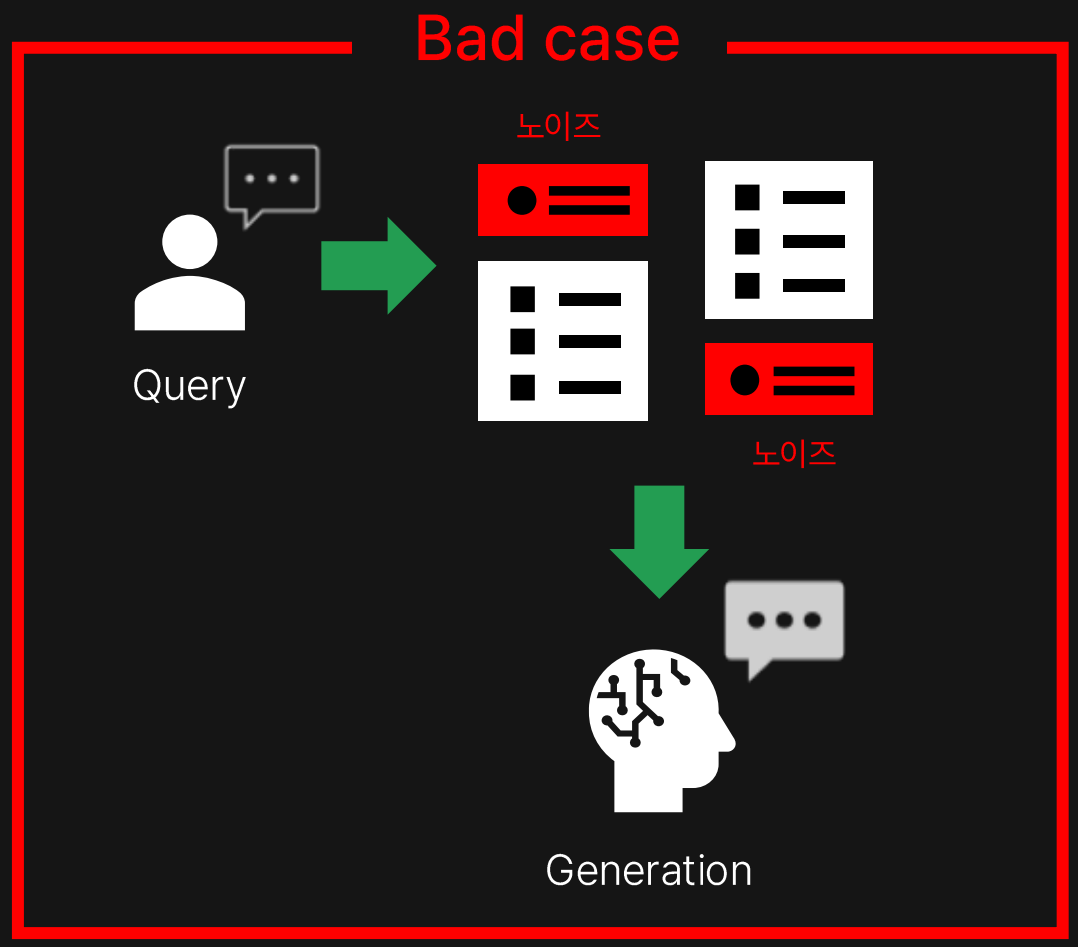
Query Routing
Query Rewriting
Query Expansion

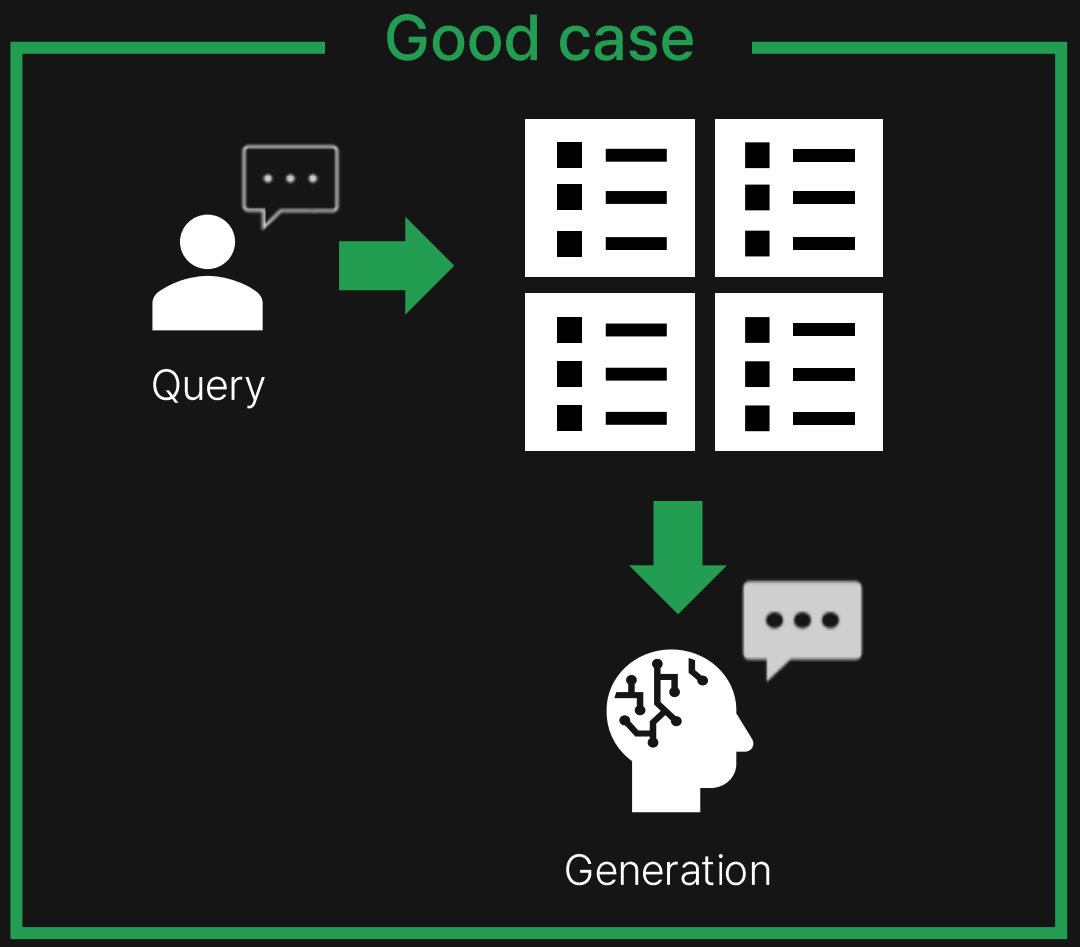
Rerank
Summary
Fusion

데이터를 일반화하기 어려움

데이터가 병목이 되는 이유 2

만약 데이터 파트를 제대로 하지 않는다면





PART 02

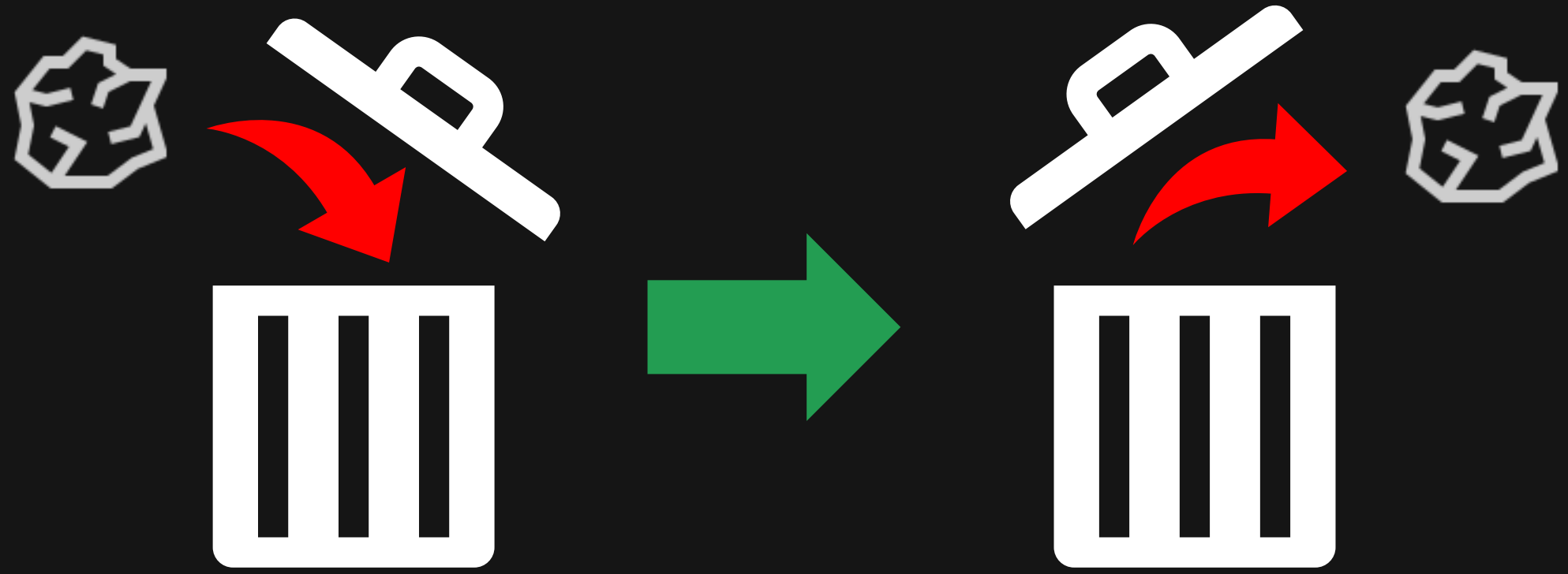
Document Structuring

"구조화는 AI가 잘 동작할 수 있는 상태로 만드는 과정이다."

02

Garbage In, Garbage Out

잘 되지 않는 대부분의 이유는 데이터

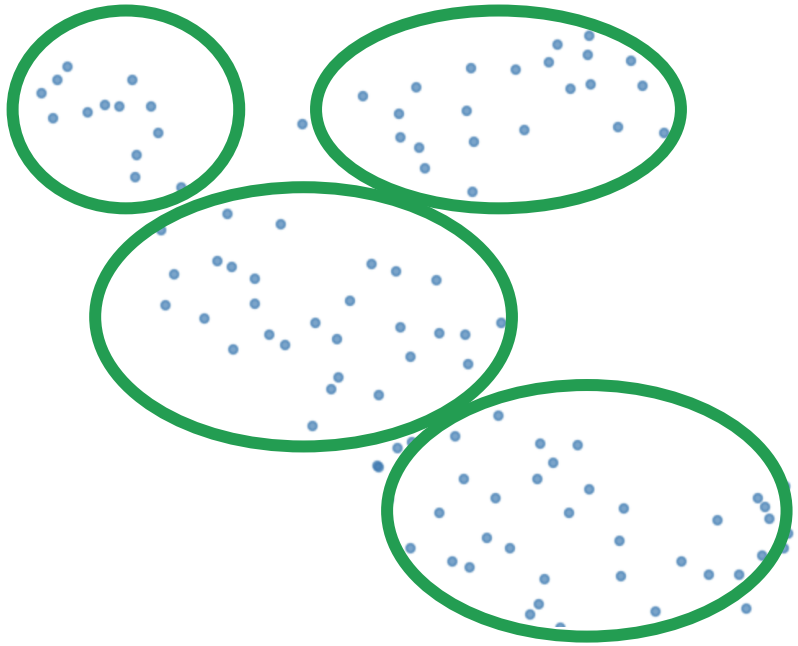


따라서, 데이터를 잘 넣는 게 중요

데이터가 잘 들어간 경우와 아닌 경우

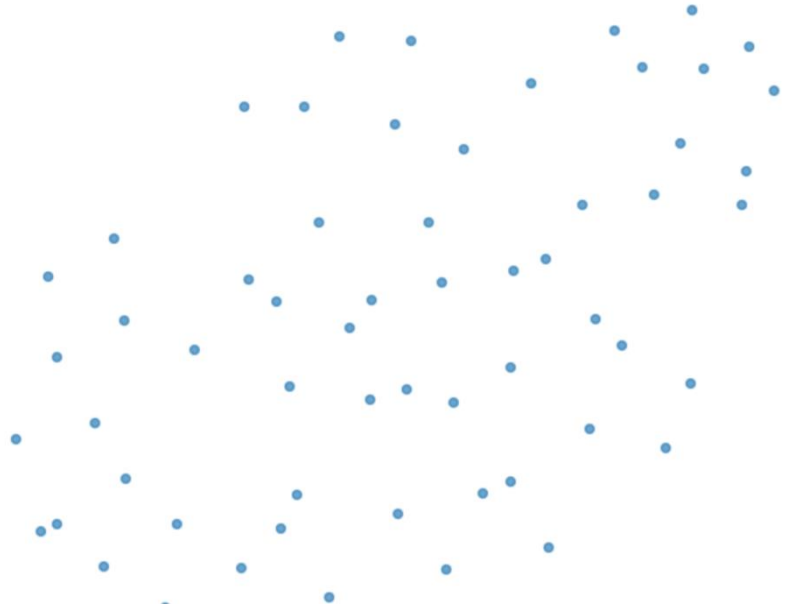
기업 문서 하나를 저장한다고 생각해보면,

Good case



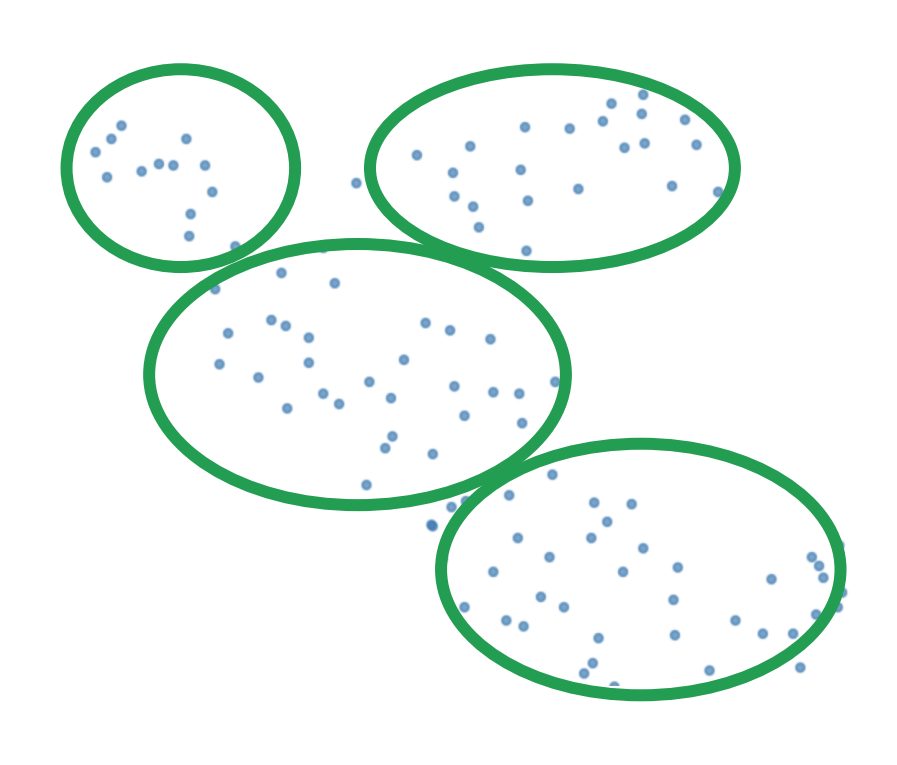
Indexing

Bad case



데이터가 잘 들어간 경우와 아닌 경우

Indexing



벡터의 군집화

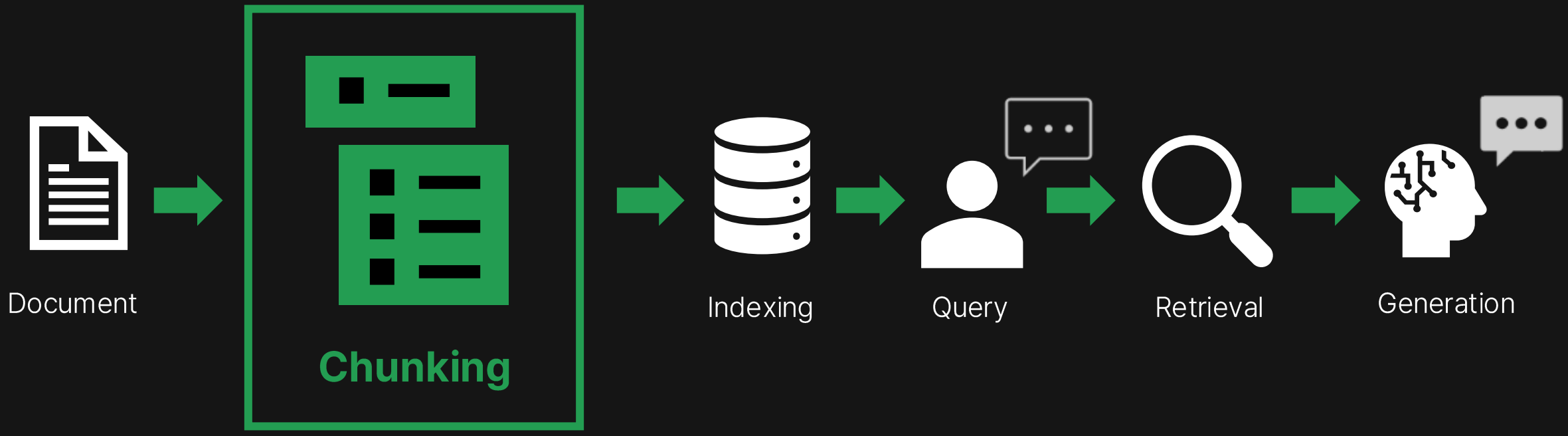
문서의 의미적 분할

질문 검색 / 답변 생성에 영향

기업 문서 수 백개를 저장한다고 하면,

데이터를 잘 넣기 위해서는?

Chunking이 중요!!



여러가지 Chunking 방법론

매뉴얼을 신입사원이 보기 좋게 만드는 것에 비유해서 생각하기

Fixed-size Chunking 고정된 크기로 Chunking하는 전략

Page-based Chunking 페이지 단위 Chunking 전략

Semantic Chunking 같은 주제의 문장/문단을 그룹화하는 Chunking 전략

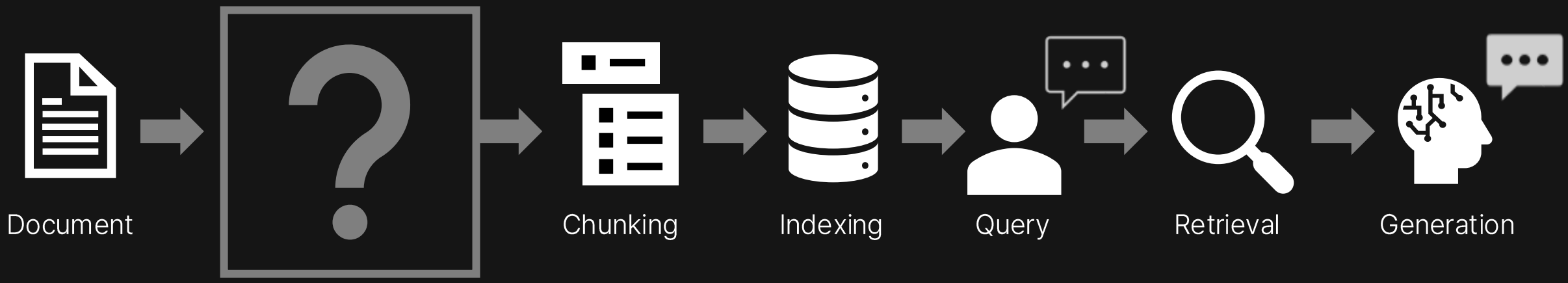
Structured Chunking Json, MD, HTML 태그 등 형식을 구조로 Chunking하는 전략

LLM-based Chunking LLM이 최적의 분할 지점을 결정하는 Chunking 전략

어떤 Chunking 방법론을 선택해야하지...

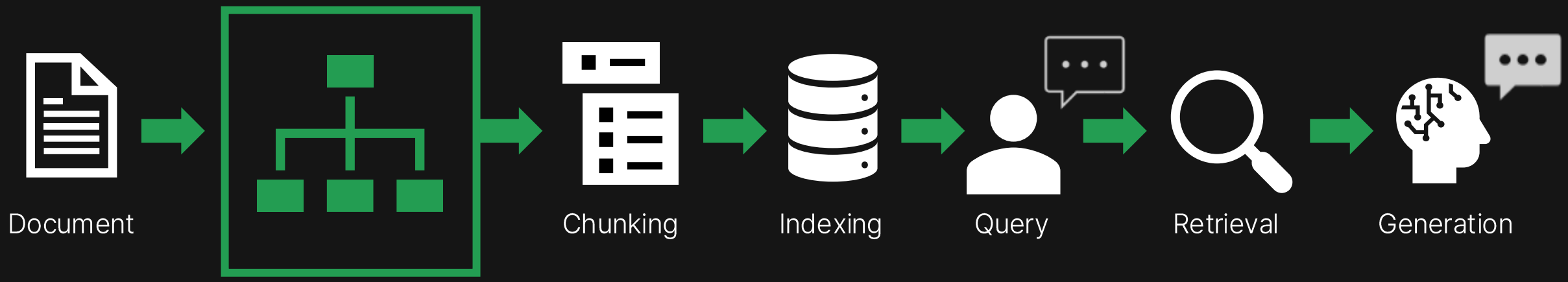
Chunking ≠ Structuring

의미 단위로 쪼개는 건 확실한데...

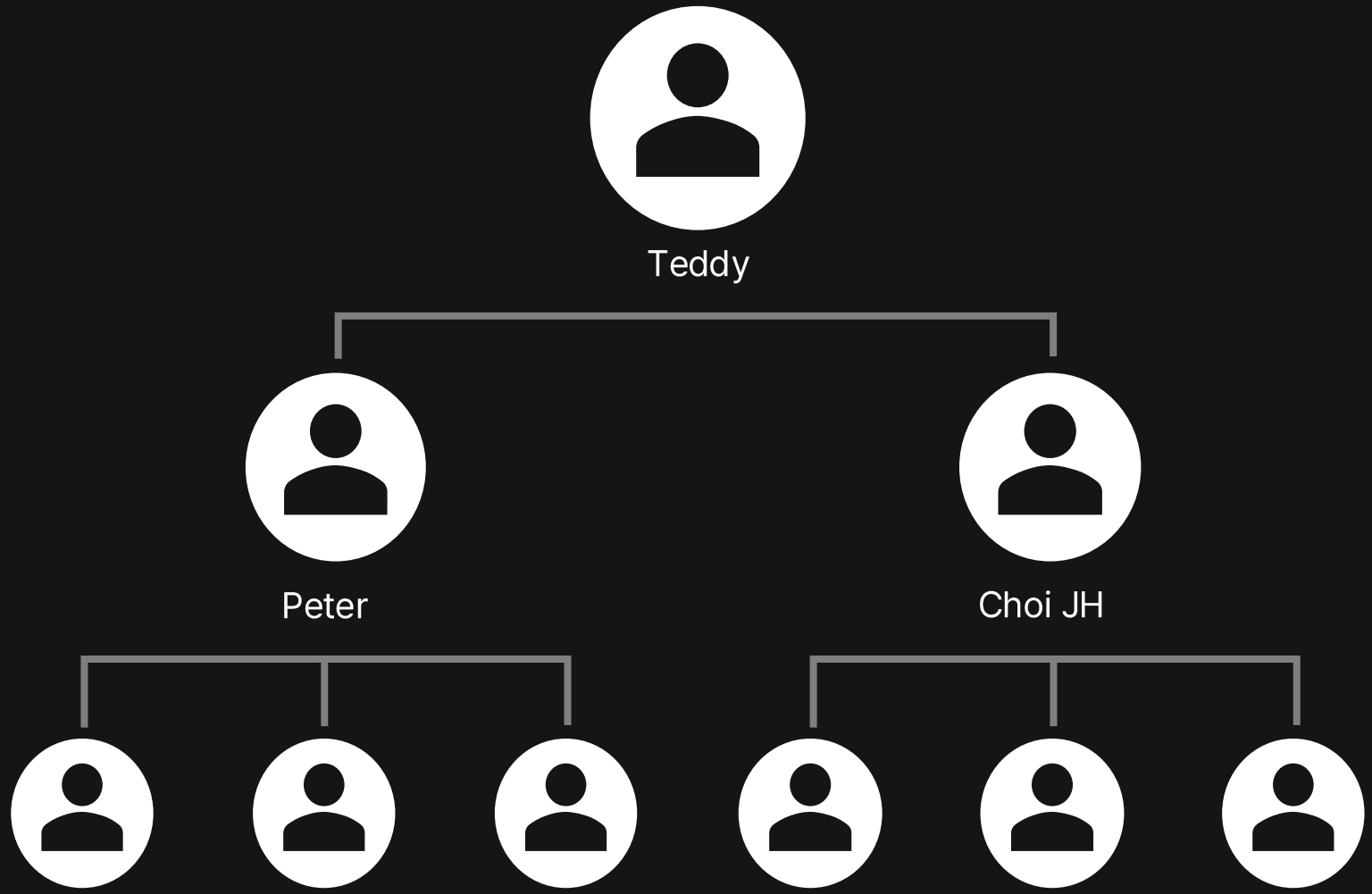


Chunking ≠ Structuring

의미 단위 보존을 위해서 Structuring 필요



Structuring 중요한 이유



Structuring 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

PART 03

Structuring 방법론

"구조화를 위한 수단"

03

Structuring 방법론 3가지

규칙 기반

명시적인 규칙으로 문서 구조를 정의하는 방식

- Heading Prefix 기반 구조 추출
- 정형 문서에서 빠르고 비용이 낮음
- 규칙 설계와 유지에 사람 개입이 필요

LLM 기반

LLM에게 문서를 해석 및 구조화 하도록 맡기는 방식

- 문맥을 이해해 유연한 구조화 가능
- 다양한 문서 포맷에 적용 가능

모델링 기반

문서 구조 자체를 학습하는 모델을 만드는 방식

- 문서 계층과 의미 단위를 직접 예측
- 대규모 문서에 안정적이고 반복 적용 가능

① 규칙 기반 접근

Heading의 규칙을 찾아 구조화

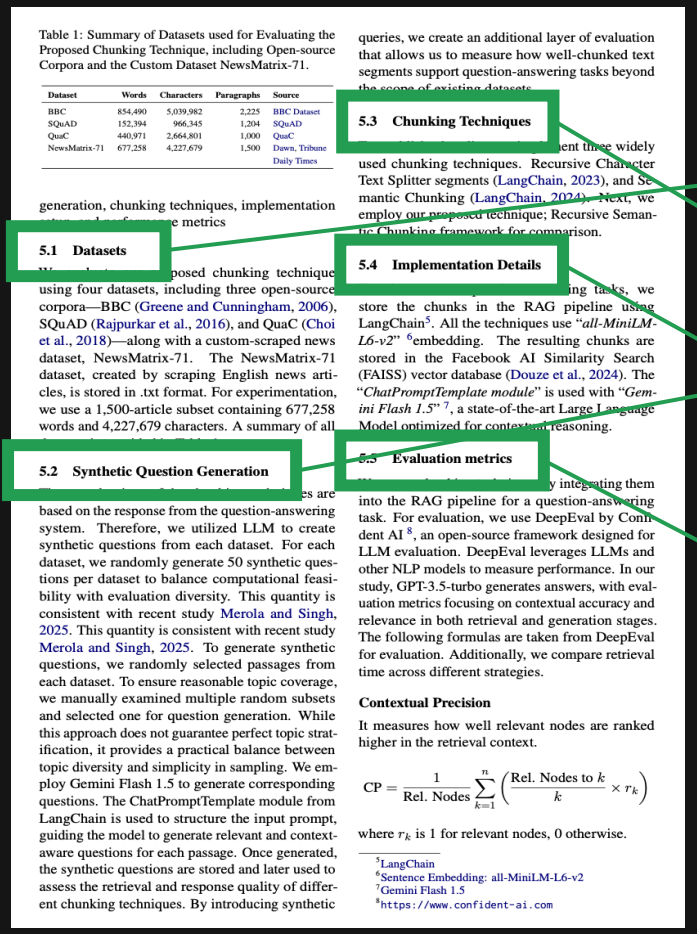


Table 1: Summary of Datasets used for Evaluating the Proposed Chunking Technique, including Open-source Corpora and the Custom Dataset NewsMatrix-71.

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 e 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

① 규칙 기반 접근의 한계

사람이 일일이 문서를 파악 및 이해 해야함

Prefix가 없는 경우도 있음

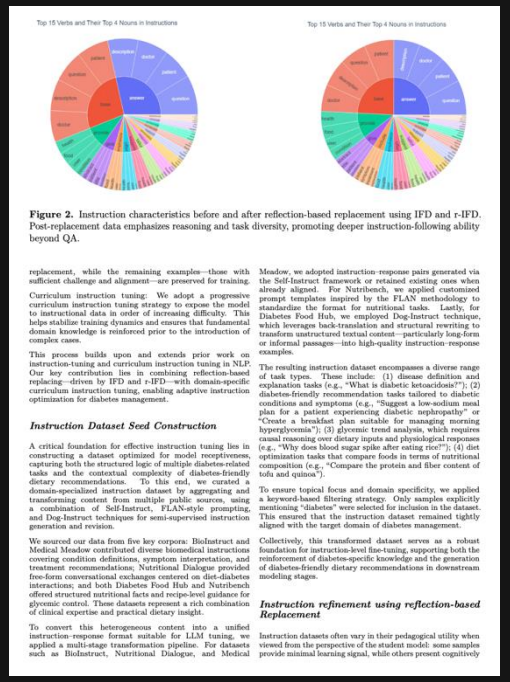
Prefix가 heading에만 있으리라는 보장이 없음

Figure 2. *Instruction Dataset Seed Construction*

Instruction refinement using reflection-based Replacement

② LLM 기반 접근 1

비용/시간 최적화 가능하면 효과적



Top 15 Verbs and Their Top 4 Nouns in Instructions

Top 15 Verbs and Their Top 4 Nouns in Instructions

Figure 2. Instruction characteristics before and after reflection-based replacement using IFD and r-IFD. Post-replacement data emphasizes reasoning and task diversity, promoting deeper instruction-following ability beyond QA.

replacement, while the remaining examples—those with sufficient challenge and alignment—are preserved for training.

Curriculum instruction tuning: We adopt a progressive curriculum instruction tuning strategy to expose the model to instructional data in order of increasing difficulty. This helps stabilize training dynamics and ensures that fundamental domain knowledge is reinforced prior to the introduction of complex cases.

This process builds upon and extends prior work on instruction-tuning and curriculum instruction tuning in NLP. Our key contribution lies in combining reflection-based replacing—driven by IFD and r-IFD—with domain-specific curriculum instruction tuning, enabling adaptive instruction optimization for diabetes management.

Instruction Dataset Seed Construction

A critical foundation for effective instruction tuning lies in constructing a dataset optimized for model receptiveness, capturing both the structured logic of multiple diabetes-related tasks and the contextual complexity of diabetes-friendly dietary recommendations. To this end, we curated a domain-specialized instruction dataset by aggregating and transforming content from multiple public sources, using a combination of Self-Instruct, FLAN-style prompting, and Dog-Instruct techniques for semi-supervised instruction generation and revision.

We sourced our data from five key corpora: BioInstruct and Medical Meadow contributed diverse biomedical instructions covering condition definitions, symptom interpretation, and treatment recommendations; Nutritional Dialogue provided free-form conversational exchanges centered on diet-diabetes interactions; and both Diabetes Food Hub and Nutribench offered structured nutritional facts and recipe-level guidance for glycemic control. These datasets represent a rich combination of clinical expertise and practical dietary insight.

To convert this heterogeneous content into a unified instruction–response format suitable for LLM tuning, we applied a multi-stage transformation pipeline. For datasets such as BioInstruct, Nutritional Dialogue, and Medical Meadow, we adopted instruction–response pairs generated via the Self-Instruct framework or retained existing ones when already aligned. For Nutribench, we applied customized prompt templates inspired by the FLAN methodology to standardize the format for nutritional tasks. Lastly, for Diabetes Food Hub, we employed Dog-Instruct technique, which leverages back-translation and structural rewriting to transform unstructured textual content—particularly long-form or informal passages—into high-quality instruction–response examples.

The resulting instruction dataset encompasses a diverse range of task types. These include: (1) disease definition and explanation tasks (e.g., "What is diabetic ketoacidosis?"); (2) diabetes-friendly recommendation tasks tailored to diabetic conditions and symptoms (e.g., "Suggest a low-sodium meal plan for a patient experiencing diabetic nephropathy" or "Create a breakfast plan suitable for managing morning hyperglycemia"); (3) glycemic trend analysis, which requires causal reasoning over dietary inputs and physiological responses (e.g., "Why does blood sugar spike after eating rice?"); (4) diet optimization tasks that compare foods in terms of nutritional composition (e.g., "Compare the protein and fiber content of tofu and quinoa").

To ensure topical focus and domain specificity, we applied a keyword-based filtering strategy. Only samples explicitly mentioning "diabetes" were selected for inclusion in the dataset. This ensured that the instruction dataset remained tightly aligned with the target domain of diabetes management.

Collectively, this transformed dataset serves as a robust foundation for instruction-level fine-tuning, supporting both the reinforcement of diabetes-specific knowledge and the generation of diabetes-friendly dietary recommendations in downstream modeling stages.

Instruction refinement using reflection-based Replacement

Instruction datasets often vary in their pedagogical utility when viewed from the perspective of the student model: some samples provide minimal learning signal, while others present cognitively

"문서를 구조화해줘"

Instruction Dataset Seed Construction
A critical foundation for effective instruction tuning lies in constructing
receptiveness, capturing both the structured logic of multiple diabetes-rela
complexity of diabetes-friendly dietary recommendations. To this end, we cur
dataset by aggregating transformation-oriented instructions from multiple pu
CoT-Structured, FLAN-style prompting, and DoG-Instruct techniques for self-s
and revision.

We sourced our data from five key corpora: BioInstruct and Medical Meadow co
instruction covering condition definitions, symptom interpretation, and trea
Nutrition Dialogue provided free-form conversational exchanges centered on d
both Diabetes Food Hub and Nutrition Encyclopedia supplied structured nutrit
knowledge, offering a balance of clinical expertise and practical dietary in

To convert this heterogeneous content into a unified instruction-response fo
we applied a multi-stage transformation pipeline. For datasets such as BioIn
Medical Meadow, we adopted instruction-response pairs generated via the Self
existing ones where already aligned. For Nutrition Encyclopedia and Diabetes
prompt templates inspired by the FLAN methodology to transform raw factual c
pairs.

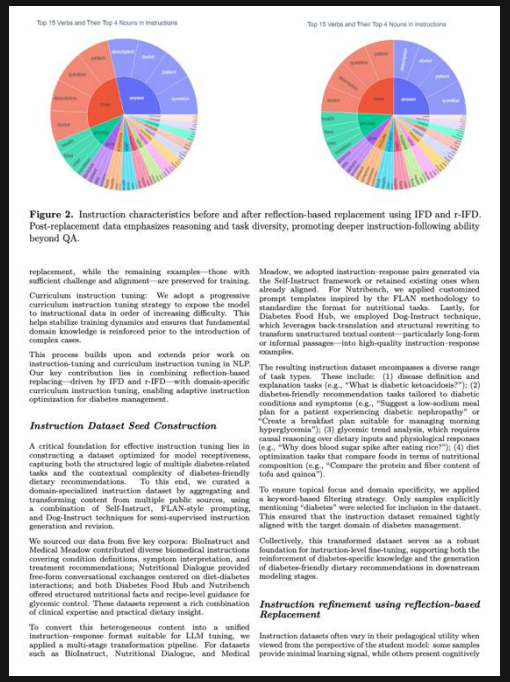
These transformed datasets were further refined through task-specific filter
reasoning, explanation, and decision-making tasks. The resulting instruction
range of task types. These include: (1) disease definition and explanation t
ketoacidosis?"), (2) diabetes education tasks related to diabetic conditions
low-sodium meal plan for a patient experiencing gestational diabetes"), (3)
(e.g., "Create a breakfast plan suitable for managing blood glucose levels")
nutritional analysis tasks (e.g., "Compare the protein and fiber content of

To ensure topical focus and domain specificity, we applied a keyword-based f
explicitly mentioning "diabetes" were selected. This ensured that the instru
aligned with the target domain of diabetes management. Collectively, this tr
robust foundation for instruction-level fine-tuning, supporting the reinforc
knowledge and the generation of diabetes-friendly dietary recommendations in

Instruction refinement using reflection-based Replacement
Instruction datasets often vary in their pedagogical utility when viewed fro
model: some samples provide minimal learning signal, while others present co

② LLM 기반 접근 2

비용/시간 최적화 가능하면 효과적



Top 15 Verbs and Their Top 4 Nouns in Instructions

Top 15 Verbs and Their Top 4 Nouns in Instructions

Figure 2. Instruction characteristics before and after reflection-based replacement using IFD and r-IFD. Post-replacement data emphasizes reasoning and task diversity, promoting deeper instruction-following ability beyond QA.

replacement, while the remaining examples—those with sufficient challenge and alignment—are preserved for training.

Curriculum instruction tuning: We adopt a progressive curriculum instruction tuning strategy to expose the model to instructional data in order of increasing difficulty. This helps stabilize training dynamics and ensures that fundamental domain knowledge is reinforced prior to the introduction of complex cases.

This process builds upon and extends prior work on instruction-tuning and curriculum instruction tuning in NLP. Our key contribution lies in combining reflection-based replacing—driven by IFD and r-IFD—with domain-specific curriculum instruction tuning, enabling adaptive instruction optimization for diabetes management.

Instruction Dataset Seed Construction

A critical foundation for effective instruction tuning lies in constructing a dataset optimized for model receptiveness, capturing both the structured logic of multiple diabetes-related tasks and the contextual complexity of diabetes-friendly dietary recommendations. To this end, we curated a domain-specialized instruction dataset by aggregating and transforming content from multiple public sources, using a combination of Self-Instruct, FLAN-style prompting, and Dog-Instruct techniques for semi-supervised instruction generation and revision.

We sourced our data from five key corpora: BioInstruct and Medical Meadow contributed diverse biomedical instructions covering condition definitions, symptom interpretation, and treatment recommendations; Nutritional Dialogue provided free-form conversational exchanges centered on diet-diabetes interactions; and both Diabetes Food Hub and Nutribench offered structured nutritional facts and recipe-level guidance for glycemic control. These datasets represent a rich combination of clinical expertise and practical dietary insight.

To convert this heterogeneous content into a unified instruction–response format suitable for LLM tuning, we applied a multi-stage transformation pipeline. For datasets such as BioInstruct, Nutritional Dialogue, and Medical Meadow, we adopted instruction–response pairs generated via the Self-Instruct framework or retained existing ones when already aligned. For Nutribench, we applied customized prompt templates inspired by the FLAN methodology to standardize the format for nutritional tasks. Lastly, for Diabetes Food Hub, we employed Dog-Instruct technique, which leverages back-translation and structural rewriting to transform unstructured textual content—particularly long-form or informal passages—into high-quality instruction–response examples.

The resulting instruction dataset encompasses a diverse range of task types. These include: (1) disease definition and explanation tasks (e.g., "What is diabetic ketoacidosis?"); (2) diabetes-friendly recommendation tasks tailored to diabetic conditions and symptoms (e.g., "Suggest a low-sodium meal plan for a patient experiencing diabetic nephropathy" or "Create a breakfast plan suitable for managing morning hyperglycemia"); (3) glycemic trend analysis, which requires causal reasoning over dietary inputs and physiological responses (e.g., "Why does blood sugar spike after eating rice?"); (4) diet optimization tasks that compare foods in terms of nutritional composition (e.g., "Compare the protein and fiber content of tofu and quinoa").

To ensure topical focus and domain specificity, we applied a keyword-based filtering strategy. Only samples explicitly mentioning "diabetes" were selected for inclusion in the dataset. This ensured that the instruction dataset remained tightly aligned with the target domain of diabetes management.

Collectively, this transformed dataset serves as a robust foundation for instruction-level fine-tuning, supporting both the reinforcement of diabetes-specific knowledge and the generation of diabetes-friendly dietary recommendations in downstream modeling stages.

Instruction refinement using reflection-based Replacement

Instruction datasets often vary in their pedagogical utility when viewed from the perspective of the student model: some samples provide minimal learning signal, while others present cognitively

"이미지의 원문 텍스트는
재출력하지 말고,
계층 경계에 해당하는
문단의 시작 문장만
순서대로 반환해줘."

```
replacement, while the remaining examples-those with sufficient challenge an

Instruction Dataset Seed Construction

Instruction refinement using reflection-based Replacement
```

② LLM 기반 접근 한계

Figure 2. Instruction characteristics before and after reflection-based replacement using IFD and r-IFD. Post-replacement data emphasizes reasoning and task diversity, promoting deeper instruction-following ability beyond QA.

replacement, while the remaining examples—those with sufficient challenge and alignment—are preserved for training.

Curriculum instruction tuning: We adopt a progressive curriculum instruction tuning strategy to expose the model to instructional data in order of increasing difficulty. This helps stabilize training dynamics and ensures that fundamental domain knowledge is reinforced prior to the introduction of complex cases.

This process builds upon and extends prior work on instruction-tuning and curriculum instruction tuning in NLP. Our key contribution lies in combining reflection-based replacing—driven by IFD and r-IFD—with domain-specific curriculum instruction tuning, enabling adaptive instruction optimization for diabetes management.

Instruction Dataset Seed Construction

A critical foundation for effective instruction tuning lies in constructing a dataset optimized for model receptiveness, capturing both the structured logic of multiple diabetes-related tasks and the contextual complexity of diabetes-friendly dietary recommendations. To this end, we curated a domain-specialized instruction dataset by aggregating and transforming content from multiple public sources, using a combination of Self-Instruct, FLAN-style prompting, and Dog-Instruct techniques for semi-supervised instruction generation and revision.

We sourced our data from five key corpora: BioInstruct and Medical Meadow contributed diverse biomedical instructions covering condition definitions, symptom interpretation, and treatment recommendations; Nutritional Dialogue provided free-form conversational exchanges centered on diet-diabetes interactions; and both Diabetes Food Hub and Nutribench offered structured nutritional facts and recipe-level guidance for glycemic control. These datasets represent a rich combination of clinical expertise and practical dietary insight.

To convert this heterogeneous content into a unified instruction–response format suitable for LLM tuning, we applied a multi-stage transformation pipeline. For datasets such as BioInstruct, Nutritional Dialogue, and Medical

Instruction refinement using reflection-based Replacement

Instruction datasets often vary in their pedagogical utility when viewed from the perspective of the student model: some samples provide minimal learning signal, while others present cognitively

"문서를 구조화해줘"

"이미지의 원문 텍스트는
재출력하지 말고,
계층 경계에 해당하는
문단의 시작 문장만
순서대로 반환해줘."

시간과 비용이 발생

사람이 직접 프롬프트 작성해야 함

③ Modeling 기반 접근

데이터

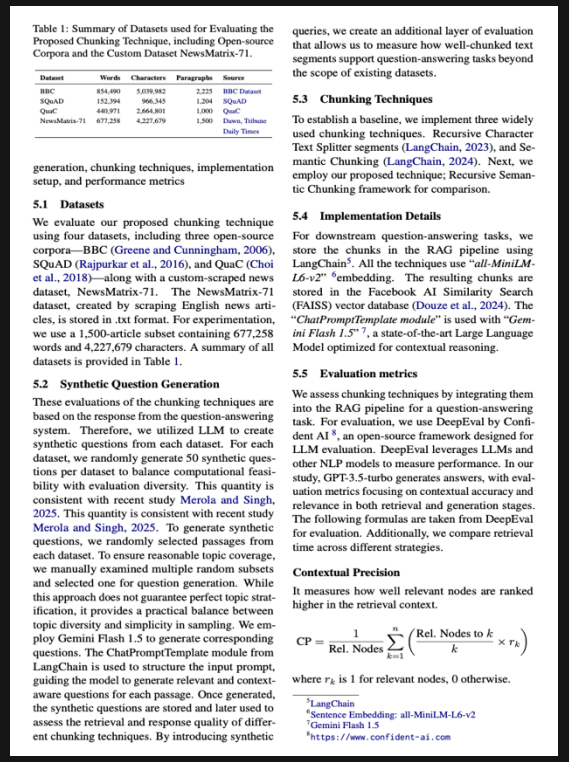


Table 1: Summary of Datasets used for Evaluating the Proposed Chunking Technique, including Open-source Corpora and the Custom Dataset NewsMatrix-71.

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

모델

결과

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

③ Modeling 기반 접근 (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

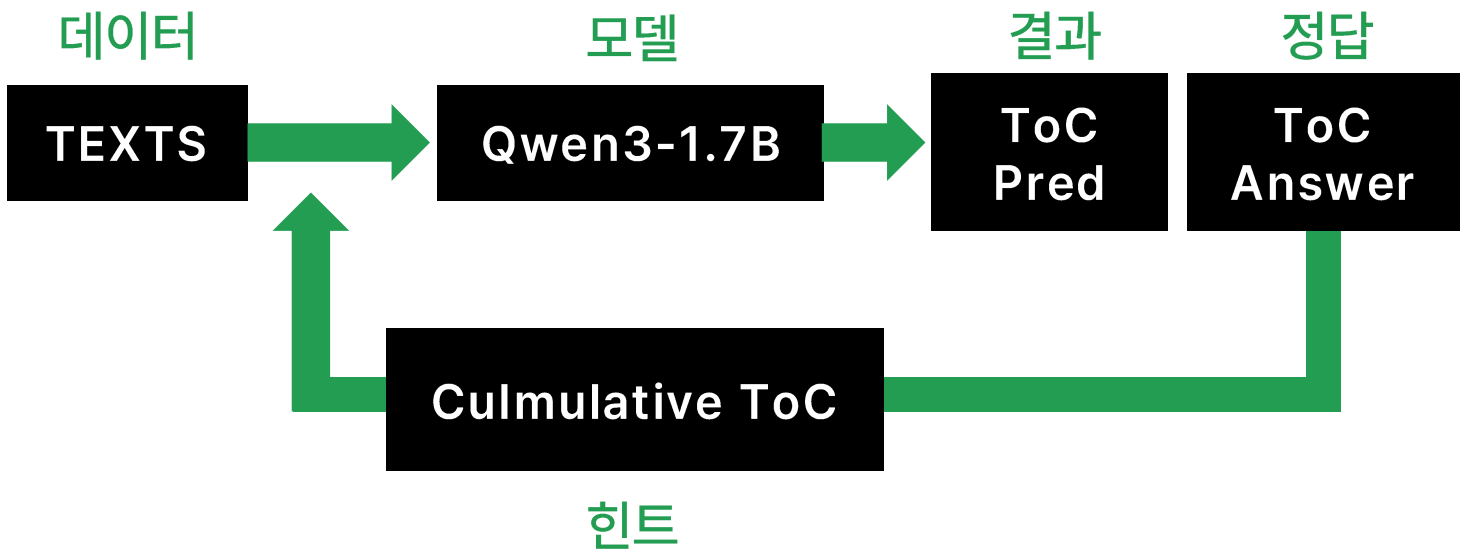
Diabetes is one of the most prevalent and complex chronic diseases globally, currently affecting more than 500 million individuals and contributing significantly to the global burden of non-communicable diseases. According to a comprehensive pooled analysis of 1108 population-representative studies including 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[45]. 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



③ Modeling 기반 접근 (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

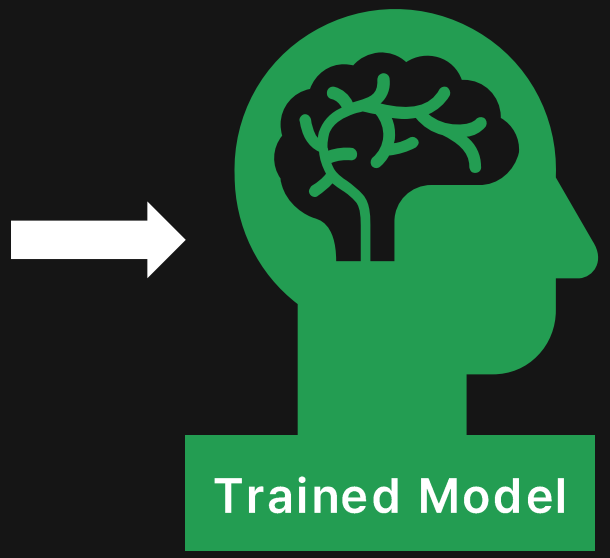
Diabetes is one of the most prevalent and complex chronic diseases globally, currently affecting more than 500 million individuals and contributing significantly to the global burden of non-communicable diseases. According to a comprehensive pooled analysis of 1108 population-representative studies including 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
```

LLM과 Modeling의 차이점

LLM

문서 덩어리를 한 번에 입력하여 Chunk 생성

- 입력 문서가 길어질 수록 할루시네이션 발생

Modeling

문서를 3문장 정도로 잘게 잘라서 입력하여 Chunk 생성

- 요청마다의 내용이 짧아서 할루시네이션이 발생 X
- LLM이 잘하는 Next Token Prediction을 더 잘할 수 있게 하는 구조
- 전역 정보도 추가로 학습해, 나무만 보지 않고, 숲도 보는 구조

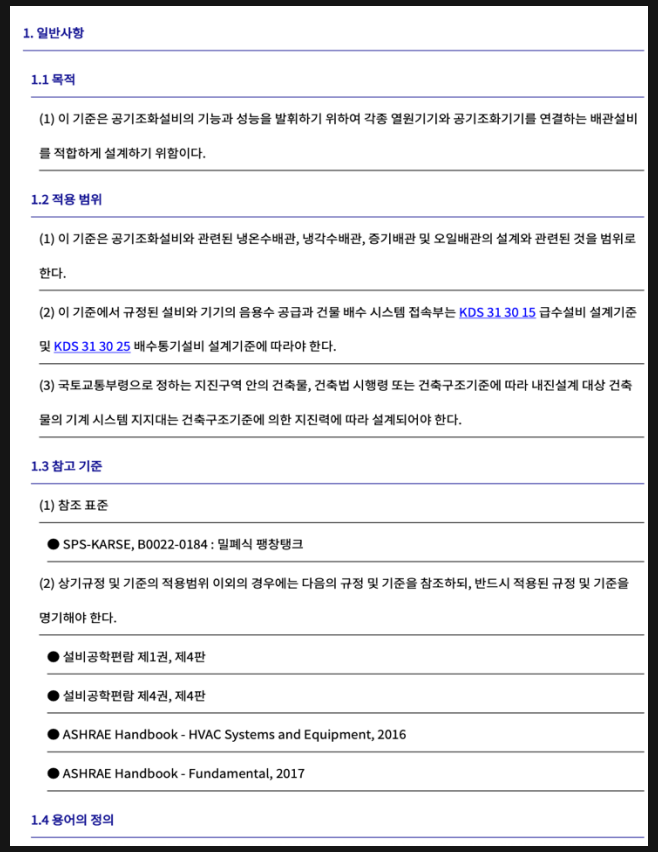
PART 04

적용 사례

"구조화된 데이터는 AI 활용 범위를 근본적으로 바꾼다."

04

건설 도메인에 대한 적용 사례



1. 일반사항

1.1 목적

(1) 이 기준은 공기조화설비의 기능과 성능을 발휘하기 위하여 각종 열원기기와 공기조화기기를 연결하는 배관설비를 적합하게 설계하기 위함이다.

1.2 적용 범위

(1) 이 기준은 공기조화설비와 관련된 냉온수배관, 냉각수배관, 증기배관 및 오일배관의 설계와 관련된 것을 범위로 한다.

(2) 이 기준에서 규정된 설비와 기기의 음용수 공급과 건물 배수 시스템 접속부는 KDS 31 30 15 급수설비 설계기준 및 KDS 31 30 25 배수통기설비 설계기준에 따라야 한다.

(3) 국토교통부령으로 정하는 지진구역 안의 건축물, 건축법 시행령 또는 건축구조기준에 따라 내진설계 대상 건축물의 기계 시스템 지지대는 건축구조기준에 의한 지진력에 따라 설계되어야 한다.

1.3 참고 기준

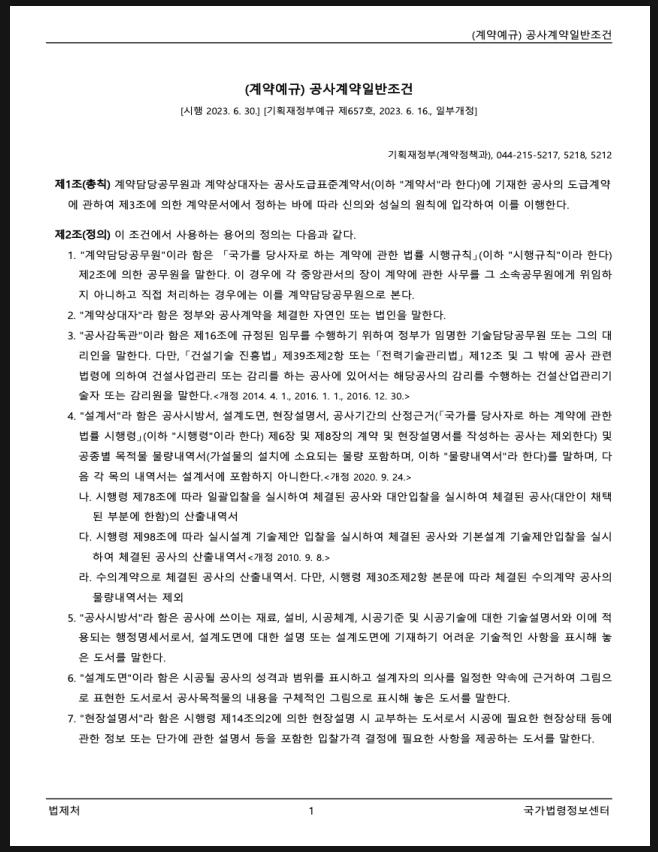
(1) 참조 표준

- SPS-KARSE, B0022-0184 : 밀폐식 팽창탱크

(2) 상기규정 및 기준의 적용범위 이외의 경우에는 다음의 규정 및 기준을 참조하되, 반드시 적용된 규정 및 기준을 명기해야 한다.

- 설비공학편람 제1권, 제4판
- 설비공학편람 제4권, 제4판
- ASHRAE Handbook - HVAC Systems and Equipment, 2016
- ASHRAE Handbook - Fundamental, 2017

1.4 용어의 정의



(계약예규) 공사계약일반조건

(계약예규) 공사계약일반조건

[시행 2023. 6. 30.] [기획재정부예규 제657호, 2023. 6. 16., 일부개정]

기획재정부(계약정책과), 044-215-5217, 5218, 5212

제1조(총칙) 계약담당공무원과 계약상대자는 공사도급표준계약서(이하 "계약서"라 한다)에 기재한 공사의 도급계약에 관하여 제3조에 의한 계약문서에서 정하는 바에 따라 신의와 성실의 원칙에 입각하여 이를 이행한다.

제2조(정의) 이 조건에서 사용하는 용어의 정의는 다음과 같다.

1. "계약담당공무원"이라 함은 「국가를 당사자로 하는 계약에 관한 법률 시행규칙」(이하 "시행규칙"이라 한다) 제2조에 의한 공무원을 말한다. 이 경우에 각 중앙관서의 장이 계약에 관한 사무를 그 소속공무원에게 위임하지 아니하고 직접 처리하는 경우에는 이를 계약담당공무원으로 본다.
2. "계약상대자"라 함은 정부와 공사계약을 체결한 자연인 또는 법인을 말한다.
3. "공사감독관"이라 함은 제16조에 규정된 임무를 수행하기 위하여 정부가 임명한 기술담당공무원 또는 그의 대리인을 말한다. 다만, 「건설기술 진흥법」 제39조제2항 또는 「전력기술관리법」 제12조 및 그 밖에 공사 관련 법령에 의하여 건설사업관리 또는 감리를 하는 공사에 있어서는 해당공사의 감리를 수행하는 건설산업관리기술자 또는 감리원을 말한다.<개정 2014. 4. 1., 2016. 1. 1., 2016. 12. 30.>
4. "설계서"라 함은 공사시방서, 설계도면, 현장설명서, 공사기간의 산정근거(「국가를 당사자로 하는 계약에 관한 법률 시행령」(이하 "시행령"이라 한다) 제6장 및 제8장의 계약 및 현장설명서를 작성하는 공사는 제외한다) 및 공종별 목적물 물량내역서(가설물의 설치에 소요되는 물량 포함하며, 이하 "물량내역서"라 한다)를 말하며, 다음 각 목의 내역서는 설계서에 포함하지 아니한다.<개정 2020. 9. 24.>
   나. 시행령 제78조에 따라 일괄입찰을 실시하여 체결된 공사와 대안입찰을 실시하여 체결된 공사(대안이 채택된 부분에 한함)의 산출내역서
   다. 시행령 제98조에 따라 실시설계 기술제안 입찰을 실시하여 체결된 공사와 기본설계 기술제안입찰을 실시하여 체결된 공사의 산출내역서<개정 2010. 9. 8.>
   라. 수의계약으로 체결된 공사의 산출내역서. 다만, 시행령 제30조제2항 본문에 따라 체결된 수의계약 공사의 물량내역서는 제외
5. "공사시방서"라 함은 공사에 쓰이는 재료, 설비, 시공체계, 시공기준 및 시공기술에 대한 기술설명서와 이에 적용되는 행정명세서로서, 설계도면에 대한 설명 또는 설계도면에 기재하기 어려운 기술적인 사항을 표시해 놓은 도서를 말한다.
6. "설계도면"이라 함은 시공될 공사의 성격과 범위를 표시하고 설계자의 의사를 일정한 약속에 근거하여 그림으로 표현한 도서로서 공사목적물의 내용을 구체적인 그림으로 표시해 놓은 도서를 말한다.
7. "현장설명서"라 함은 시행령 제14조의2에 의한 현장설명 시 교부하는 도서로서 시공에 필요한 현장상태 등에 관한 정보 또는 단가에 관한 설명서 등을 포함한 입찰가격 결정에 필요한 사항을 제공하는 도서를 말한다.

법제처 1 국가법령정보센터

건설 도메인 설비 공종 RAG Project

- QA

Challenges

- 건설 용어
- 수식/도면
- 다양한 형태의 문서

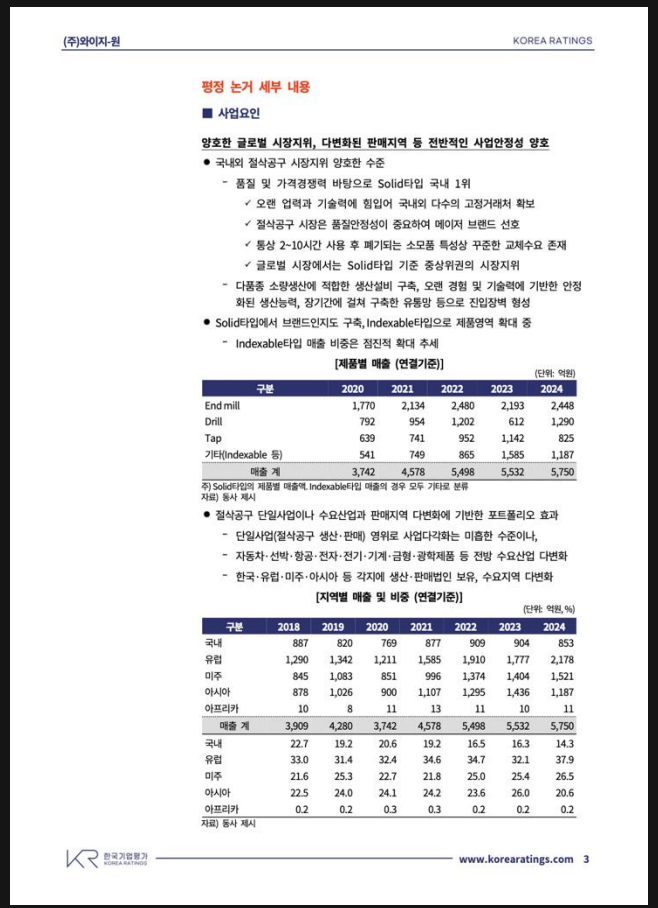
Results

- Structuring + Chunking만으로 **85% Recall**

시사점

- 데이터에 대한 이해와 구조화가 핵심 기법

금융 도메인에 대한 적용 사례



(주)와이지-원 KOREA RATINGS

평정 논거 세부 내용

■ 사업요인

양호한 글로벌 시장지위, 다변화된 판매지역 등 전반적인 사업안정성 양호

- 국내외 절삭공구 시장지위 양호한 수준
  - 품질 및 가격경쟁력 바탕으로 Solid타입 국내 1위
    - ✓ 오랜 업력과 기술력에 힘입어 국내외 다수의 고정거래처 확보
    - ✓ 절삭공구 시장은 품질안정성이 중요하여 메이저 브랜드 선호
    - ✓ 통상 2~10시간 사용 후 폐기되는 소모품 특성상 꾸준한 교체수요 존재
    - ✓ 글로벌 시장에서는 Solid타입 기준 중상위권의 시장지위
  - 다품종 소량생산에 적합한 생산설비 구축, 오랜 경험 및 기술력에 기반한 안정화된 생산능력, 장기간에 걸쳐 구축한 유통망 등으로 진입장벽 형성
- Solid타입에서 브랜드인지도 구축, Indexable타입으로 제품영역 확대 중
  - Indexable타입 매출 비중은 점진적 확대 추세

[제품별 매출 (연결기준)]

(단위: 억원)

구분	2020	2021	2022	2023	2024
End mill	1,770	2,134	2,480	2,193	2,448
Drill	792	954	1,202	612	1,290
Tap	639	741	952	1,142	825
기타(Indexable 등)	541	749	865	1,585	1,187
매출 계	3,742	4,578	5,498	5,532	5,750

주) Solid타입의 제품별 매출액. Indexable타입 매출의 경우 모두 기타로 분류
자료) 동사 제시

- 절삭공구 단일사업이나 수요산업과 판매지역 다변화에 기반한 포트폴리오 효과
  - 단일사업(절삭공구 생산·판매) 영위로 사업다각화는 미흡한 수준이나,
  - 자동차·선박·항공·전자·전기·기계·금형·광학제품 등 전방 수요산업 다변화
  - 한국·유럽·미주·아시아 등 각지에 생산·판매법인 보유, 수요지역 다변화

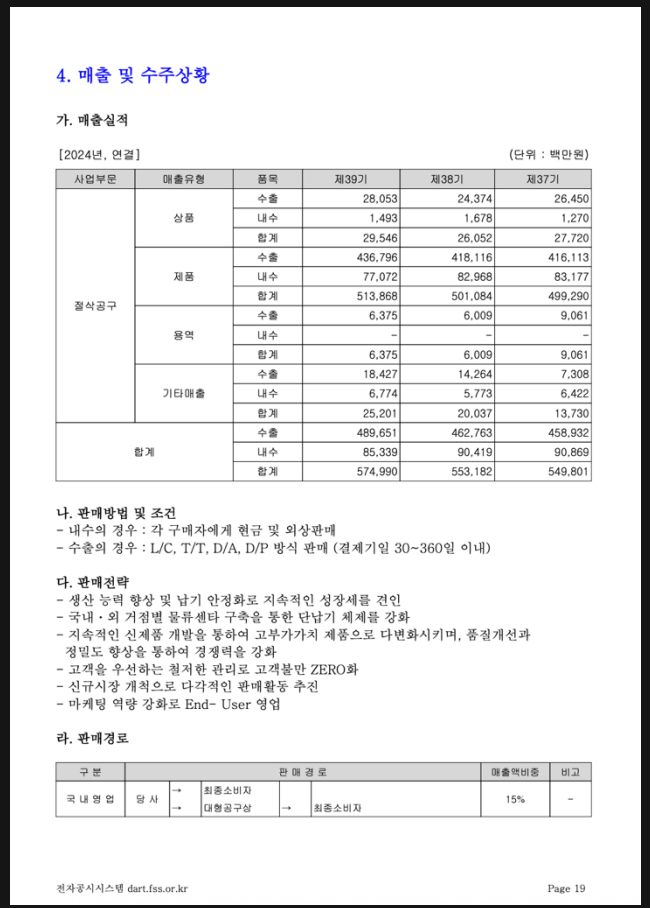
[지역별 매출 및 비중 (연결기준)]

(단위: 억원,%)

구분	2018	2019	2020	2021	2022	2023	2024
국내	887	820	769	877	909	904	853
유럽	1,290	1,342	1,211	1,585	1,910	1,777	2,178
미주	845	1,083	851	996	1,374	1,404	1,521
아시아	878	1,026	900	1,107	1,295	1,436	1,187
아프리카	10	8	11	13	11	10	11
매출 계	3,909	4,280	3,742	4,578	5,498	5,532	5,750
국내	22.7	19.2	20.6	19.2	16.5	16.3	14.3
유럽	33.0	31.4	32.4	34.6	34.7	32.1	37.9
미주	21.6	25.3	22.7	21.8	25.0	25.4	26.5
아시아	22.5	24.0	24.1	24.2	23.6	26.0	20.6
아프리카	0.2	0.2	0.3	0.3	0.2	0.2	0.2

자료) 동사 제시

한국기업평가 KOREA RATINGS www.korearatings.com 3



4. 매출 및 수주상황

가. 매출실적

[2024년, 연결] (단위 : 백만원)

사업부문	매출유형	품목	제39기	제38기	제37기
절삭공구	상품	수출	28,053	24,374	26,450
		내수	1,493	1,678	1,270
		합계	29,546	26,052	27,720
	제품	수출	436,796	418,116	416,113
		내수	77,072	82,968	83,177
		합계	513,868	501,084	499,290
	용역	수출	6,375	6,009	9,061
		내수	-	-	-
		합계	6,375	6,009	9,061
	기타매출	수출	18,427	14,264	7,308
		내수	6,774	5,773	6,422
		합계	25,201	20,037	13,730
합계		수출	489,651	462,763	458,932
		내수	85,339	90,419	90,869
		합계	574,990	553,182	549,801

나. 판매방법 및 조건

- 내수의 경우 : 각 구매자에게 현금 및 외상판매
- 수출의 경우 : L/C, T/T, D/A, D/P 방식 판매 (결제기일 30~360일 이내)

다. 판매전략

- 생산 능력 향상 및 납기 안정화로 지속적인 성장세를 견인
- 국내·외 거점별 물류센타 구축을 통한 단납기 체제를 강화
- 지속적인 신제품 개발을 통하여 고부가가치 제품으로 다변화시키며, 품질개선과 정밀도 향상을 통하여 경쟁력을 강화
- 고객을 우선하는 철저한 관리로 고객불만 ZERO화
- 신규시장 개척으로 다각적인 판매활동 추진
- 마케팅 역량 강화로 End- User 영업

라. 판매경로

구 분	판 매 경 로					매출액비중	비고
국 내 영 업	당 사	→	최종소비자			15%	-
		→	대형공구상	→	최종소비자		

전자공시시스템 dart.fss.or.kr Page 19

금융 도메인 RAG Project

- 금융 승인신청서 작성

Challenges

- 동적 문서 처리
- 새로운 문서에 대한 extraction 및 labeling
  ㄴ 의미 단위 Chunking 중요

Results

- **실무자 만족도 80% 이상**

시사점

- RAG 기법인 Metadata filtering을 위해서도 의미 단위 Chunking이 필수

결론

문서 구조화는 전처리가 아니라 AI가 문서를 이해하도록 만드는 설계

1. AX 실패의 진짜 원인

AX 실패의 가장 큰 원인은 모델이 아니라 **데이터**다.
특히 기업 문서는 구조가 무너진 상태로 RAG에 투입되고 있다.

2. Chunking의 한계

Chunking만으로는 문서를 이해할 수 없다.
의미 단위와 **문서 계층이 보존**되지 않으면 검색은 실패한다.

결론

문서 구조화는 전처리가 아니라 AI가 문서를 이해하도록 만드는 설계

3. 기업 문서의 본질

기업 문서는 본질적으로 구조적 특징을 가진 데이터다.
이 구조를 활용해 **의미 단위로 문서를 분할**하고 검색해야 한다.

4. 기업 환경에서 RAG를 잘하는 방법

RAG를 잘하기 위해서는 순서가 중요하다.
RAG 적용 이전에 문서 구조에 대한 **이해와 분석이 선행**되어야 한다.

Thank You