KoViDoRe: A Benchmark for Korean Visual Document Retrieval

Yongbin Choi, Yongwoo Song, Mujeen Sung*
Kyung Hee University
{yongbinchoi, syw5141, mujeensung}@khu.ac.kr

Code: https://github.com/whybe-choi/kovidore-benchmark
Dataset: https://hf.co/datasets/NomaDamas/ko-vdr-train-public

Abstract

Recent advances in multimodal retrieval have improved the ability to retrieve information from visually rich documents such as PDFs and reports. However, existing benchmarks remain largely centered on English and provide limited coverage of Korean visual documents with complex structures. Furthermore, most existing Korean resources primarily evaluate single-page retrieval, failing to capture realistic scenarios that require evidence aggregation across multiple pages. To address these gaps, we introduce **KoViDoRe**, a benchmark for Korean visual document retrieval. The dataset is constructed from publicly available Korean documents with diverse layouts, including tables, figures, and multi-column structures. We develop a multi-stage data curation pipeline consisting of structured document parsing, synthetic query generation using both summary-based and context-based strategies, and relevance mapping with human verification. Using KoViDoRe, we evaluate a wide range of multimodal retrieval models and observe that current models struggle to effectively handle Korean visual document retrieval, particularly in settings involving structured content and diverse query types. Motivated by this finding, we further curate a large-scale training dataset, **Ko-VDR Train Public**, to support the development of retrieval models tailored to Korean visual documents. Together, KoViDoRe and Ko-VDR Train Public provide a unified benchmark and training resource for Korean visual document retrieval.

1 Introduction

Recent advances in multimodal large language models and retrieval-augmented generation (RAG) have significantly improved the ability to retrieve and reason over complex documents (Abootorabi et al., 2025; Song et al., 2025; Yu et al., 2024). In particular, a growing line of work on visual document retrieval (VDR) and multimodal retrieval

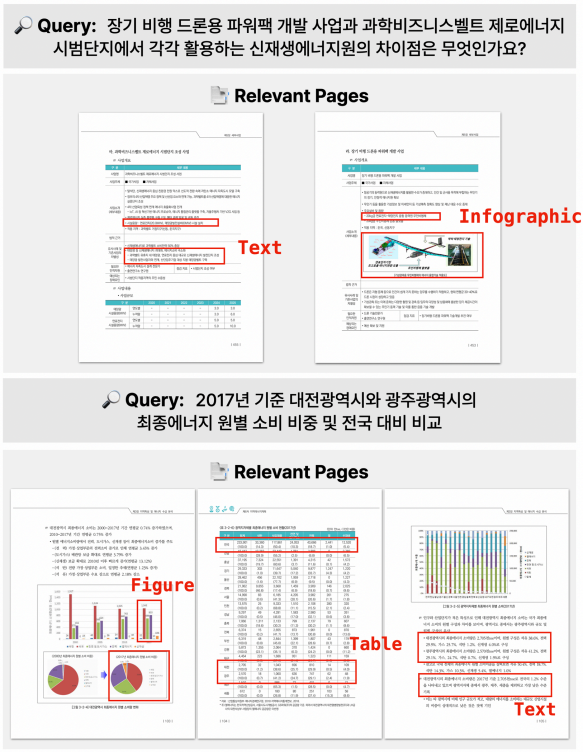


Figure 1: Examples of queries in KoViDoRe and their corresponding relevant pages. Each query requires aggregating evidence from pages and diverse modalities, including text, tables, figures, and infographics.

models, including approaches such as ColPali (Faysse et al., 2024), has demonstrated strong performance in retrieving document pages by jointly modeling textual, visual, and layout information. These developments have enabled systems to move beyond text-only retrieval and better handle structured documents such as PDFs, reports, and forms (Yan et al., 2026). To support this progress, recent benchmarks have adopted large-scale synthetic data generation pipelines, exemplified by frameworks such as ViDoRe (Macé et al., 2025; Loison et al., 2026), enabling scalable evaluation of multimodal retrieval systems. In parallel, efforts such as Jina-VDR (Günther et al., 2025), MIRACL-VISION

*Corresponding author

Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval, pages 54–80

(Osmulski et al., 2025) and SDS KoPub VDR (Lee et al., 2025) have extended this paradigm to non-English settings, providing valuable resources for Korean document retrieval and highlighting the importance of multilingual evaluation.

Despite these advances, existing benchmarks largely formulate VDR as a single-page retrieval task, where each page is treated as an independent unit (Wasserman et al., 2025; Wang et al., 2025). While this formulation simplifies evaluation, it does not accurately reflect how information is organized in real-world documents. In practical scenarios such as financial reports, policy documents, and technical manuals, relevant information is often distributed across multiple pages, requiring systems to aggregate evidence and perform reasoning over a set of pages rather than retrieving a single relevant page (Cho et al., 2024). Although prior datasets may include queries that involve reasoning, such reasoning is typically confined to a single page or limited contextual scopes (Dong et al., 2025). Addressing this limitation requires a shift from single-page retrieval to multi-page evidence aggregation, where retrieval systems must identify a coherent set of pages that collectively fulfill the information need.

In this work, we introduce **KoViDoRe**, a benchmark for Korean visual document retrieval that explicitly models this setting. Building upon prior synthetic data generation approaches, we construct a dataset of realistic enterprise-style documents and generate queries that require retrieving and synthesizing information distributed across multiple pages to provide a complete answer. Unlike existing benchmarks that primarily focus on single-page retrieval with extractive queries, we formulate the task to address multi-page relevance. In this setting, a single query often corresponds to multiple supporting pages, requiring models to identify the full set of pages whose combined content is necessary to satisfy the information need. To enable scalable construction of such queries, we adopt and adapt synthetic query generation techniques to the Korean document domain. Our pipeline leverages large language models to generate queries with diverse reasoning patterns, including multi-hop inference, numerical comparison, and cross-sectional aggregation, while maintaining explicit mappings between queries and their supporting pages. Rather than introducing a fundamentally new generation method, our focus is on restructuring the task and dataset to better reflect realistic retrieval scenarios, particularly in non-English and enterprise contexts.

In addition, we release **Ko-VDR Train Public**, a large-scale training dataset aligned with the proposed task, providing a foundation for developing and evaluating retrieval models in Korean multimodal settings. Through extensive experiments, we show that existing multimodal retrieval approaches struggle significantly under this formulation, especially as the number of required supporting pages increases. These results highlight the limitations of current single-page retrieval paradigms and underscore the need for models that can effectively aggregate over evidence distributed across multiple pages. Our contributions are as follows:

- We introduce **KoViDoRe**, a Korean-focused benchmark designed to evaluate multi-page retrieval performance in realistic document settings.
- We adapt synthetic query generation techniques to construct realistic queries with explicit page-level supervision.
- We release **Ko-VDR Train Public**, a large-scale dataset supporting training in Korean multimodal retrieval.
- We show that existing retrieval models struggle to retrieve evidence distributed across multiple pages on Korean visual documents, highlighting a gap not captured by existing benchmarks.

2 Related Work

2.1 Multimodal Retrieval Models

Recent advances in multimodal retrieval models have significantly improved the ability to retrieve information from visually rich documents (Günther et al., 2025; Ma et al., 2024; Li et al., 2026; Moreira et al., 2026). In particular, models such as ColPali and related late-interaction architectures represent each document page as a unified retrieval unit, encoding the textual content, visual features, and layout structure contained within the page (Faysse et al., 2024; Xiao et al., 2025). This allows retrieval models to capture not only semantic information from text, but also spatial and visual cues, leading to more effective retrieval over visually rich documents. These approaches have demonstrated strong performance across a variety of document understanding tasks, especially in settings where visual structure plays a critical role. In addition to

late-interaction models, dual-encoder and dense retrieval approaches have also been extended to multimodal settings, often leveraging vision-language models to capture both textual and visual semantics (Ma et al., 2024; Nomic Team, 2025). These developments have contributed to substantial progress in retrieving relevant content from structured documents such as PDFs, forms, and reports.

However, despite these modeling advances, the datasets used to train and evaluate such models are largely concentrated on English and European languages (Yu et al., 2024; Günther et al., 2025; Loison et al., 2026; Peng et al., 2025; Wasserman et al., 2025; Shorten et al., 2026). As a result, the performance and behavior of multimodal retrieval models on other languages, including Korean, remain underexplored. This is particularly important in document retrieval settings, where linguistic characteristics such as morphology, spacing variation, and domain-specific expressions interact with visual structure and layout. The lack of dedicated Korean benchmarks limits the ability to assess and develop retrieval models for realistic Korean document scenarios.

2.2 Vision Document Retrieval Benchmarks

Recent benchmarks for visual document retrieval and document-centric multimodal retrieval, such as ViDoRe (Macé et al., 2025; Loison et al., 2026), Jina-VDR (Günther et al., 2025), REAL-MM-RAG (Wasserman et al., 2025), UniDoc-Bench (Peng et al., 2025), MIRACL-VISION (Osmulski et al., 2025), and IRPAPERS (Shorten et al., 2026) have significantly expanded evaluation settings by incorporating visually rich documents, multimodal signals, and realistic query formulations. Many of these benchmarks also support multilingual evaluation. However, their language coverage is largely centered on English and European languages, leaving Korean relatively underrepresented despite its distinct linguistic and document characteristics.

Jina-VDR, for example, includes a Korean subset and broadens the diversity of visual documents and query types. However, its document collections are constructed to cover a wide range of modalities and scenarios, which can make them less representative of real-world Korean document retrieval settings, such as structured public documents, reports, or enterprise-style materials. Similarly, MIRACL-VISION provides multilingual evaluation with Korean queries and documents, but its corpus is primarily derived from Wikipedia, which differs substantially from the types of structured and visually complex documents commonly encountered in real-world Korean retrieval scenarios. This discrepancy limits its ability to fully capture the challenges of practical document retrieval in Korean contexts. SDS KoPub VDR (Lee et al., 2025) addresses this gap by introducing a large-scale benchmark for Korean visual document retrieval. It provides an important step toward evaluating retrieval models on Korean documents with complex layouts and diverse structures. Nevertheless, its query formulation remains strictly focused on single-page retrieval, where each query is mapped to only one relevant page, rather than addressing information distributed across multiple pages.

In contrast, KoViDoRe is designed to emphasize more complex information needs that cannot be satisfied by a single page alone. Our queries require aggregating evidence across multiple pages and capturing relationships between distributed pieces of information. By focusing on Korean documents while introducing queries with higher reasoning complexity and more realistic document distributions, KoViDoRe complements existing benchmarks and provides a more challenging evaluation setting for Korean visual document retrieval.

3 Dataset Curation

As illustrated in Figure 2, our dataset curation pipeline consists of document collection, structured parsing, query generation, and relevance mapping. The overall pipeline design is inspired by ViDoRe V3 (Loison et al., 2026), and is adapted to better reflect the characteristics of Korean document ecosystems and real-world data sources.

3.1 Source Collection

To construct a realistic benchmark for Korean visual document retrieval, we collect document corpora from publicly available sources, including government reports, policy documents, and enterprise-style materials obtained from the Korean public data portal[1] and official institutional websites. We prioritize documents that are freely available under the Korea Open Government License (KOGL) Type 1 and 2, as well as materials without restrictive licensing conditions. This ensures that the collected data can be used for research and downstream applications without legal constraints. Following SDS KoPub VDR (Lee et al., 2025), which

[1] https://www.data.go.kr

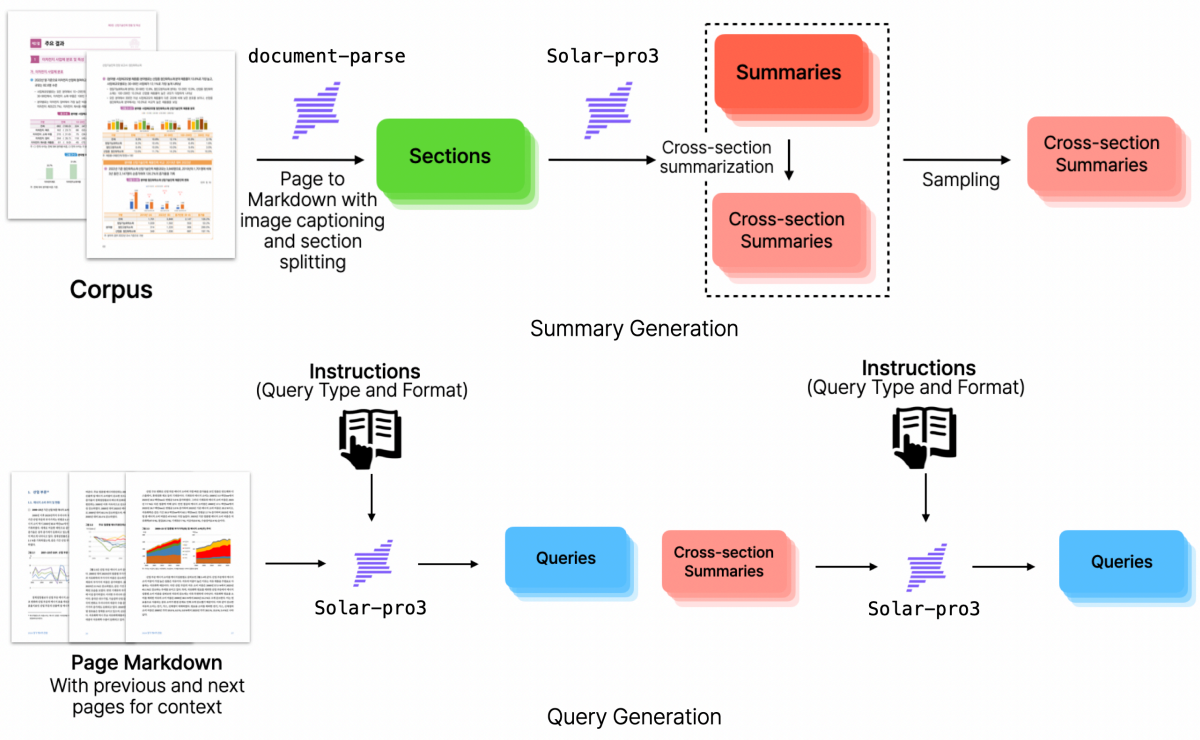


Figure 2: Overview of the KoViDoRe data generation pipeline. Queries are generated through both summary-based and context-based strategies. Summary representations capture global document relationships, while context-based generation focuses on local content. The pipeline further includes filtering and manual verification to ensure query quality and reliable relevance mapping.

also leverages publicly accessible Korean documents, we focus on authentic document sources rather than synthetic or simplified formats. This allows the benchmark to reflect real-world document characteristics, including complex layouts, tables, figures, and multi-column structures.

3.2 Document Processing and Parsing

We first split each PDF document into individual pages and perform processing at the page level. This allows us to treat each page as a fundamental unit while preserving the document structure. To extract structured representations from each page, we employ the Upstage Document Parse[2], which is effective for parsing structurally complex Korean documents while preserving layout-aware information. For every page, textual content and visual elements are identified and organized into semantically meaningful sections. Specifically, the parser provides both page-level markdown and element-level markdown for each page. The page-level markdown offers a unified view of the entire page content, while the element-level markdown decomposes the page into fine-grained components such as text blocks, tables, figures, charts, and diagrams. In addition to structural extraction, the document parser provides captions for visual elements such as figures, charts, and diagrams. These captions offer semantic descriptions of visual content, enabling better understanding of non-textual information during downstream processing. By combining page-level and element-level markdown with captioned visual elements, our preprocessing pipeline preserves document-level structure while enabling fine-grained and semantically enriched access to page content.

3.3 Query Generation

To construct a diverse and scalable set of queries, we adopt a synthetic query generation pipeline based on reasoning-oriented language model, Solar-Pro3[3], which supports strong Korean language understanding and is well-suited for generating complex Korean queries that reflect multiple information needs. The model generates queries by conditioning on document content, including both textual and visual information extracted during preprocessing.

[2]document-parse-251217

[3]solar-pro3-260126

Summary-based Query Generation Before query generation, we first construct intermediate summaries to better expose document structure and cross-page relationships. Specifically, we generate two types of summaries. The first type consists of single-section summaries that describe individual sections. The second type consists of cross-section summaries, which are constructed by randomly sampling multiple single-section summaries (e.g., 3, 5, or 7 sections) and synthesizing them to capture cross-sectional relationships. These summaries provide a higher-level abstraction of document content, allowing the generation process to capture relationships that are not easily observable from isolated pages. Queries generated from summaries therefore tend to reflect more global information needs and often require reasoning across multiple sections or pages.

Context-based Query Generation In addition to summary-based generation, we also generate queries directly from local document context. In this setting, the model is prompted using page-level or local multi-page context windows, enabling it to produce queries grounded in nearby content. This complementary route helps capture more localized information needs and preserves natural query patterns tied to specific document regions. Combining both summary-based and context-based queries allows the dataset to cover a broader spectrum of retrieval scenarios.

Diversity Control To promote diversity, we adopt the query formulation scheme introduced in ViDoRe V3 (Loison et al., 2026). Specifically, we control the query format and type during generation, enabling the construction of a diverse set of queries with different structural patterns and information needs. As a result, the dataset includes various query types such as multi-hop reasoning, numerical comparison, and aggregation-based queries, where a single query may exhibit multiple types simultaneously. The full set of query types and formats, along with their definitions, is summarized in Table 5 and Table 6.

3.4 Relevance Mapping and Filtering

To construct reliable query-document relevance annotations, we incorporate relevance mapping into multiple stages of the pipeline. During query generation, the model is provided with document content in markdown format, including both page-level and element-level representations, and queries are generated based on specific sections or combinations of sections, implicitly capturing initial relevance signals. After query generation, we perform an additional relevance mapping step at the page level by evaluating each query against candidate document pages to identify supporting evidence.

To improve annotation quality while reducing manual effort, we apply both consistency-based and rule-based filtering. Relevance signals from the generation stage are compared with those from the additional mapping stage, and only consistent pairs are retained. While this process may discard some challenging cases where relevant pages are difficult to identify during mapping, we prioritize annotation reliability over coverage, as relevance judgments can be inherently ambiguous in mul-

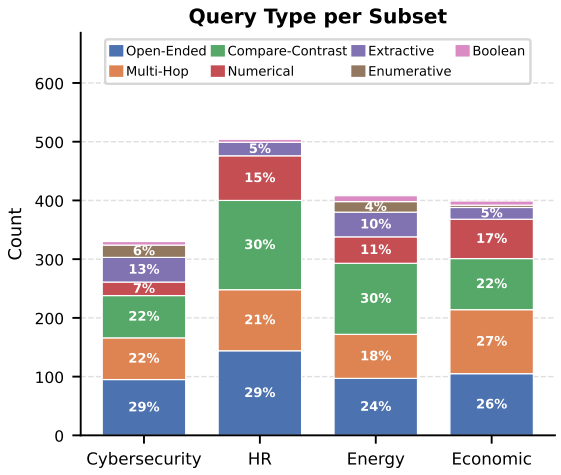


Figure 3: Distribution of query types across subsets. Note that query types are not mutually exclusive, accounting for the multi-faceted nature of complex information needs.

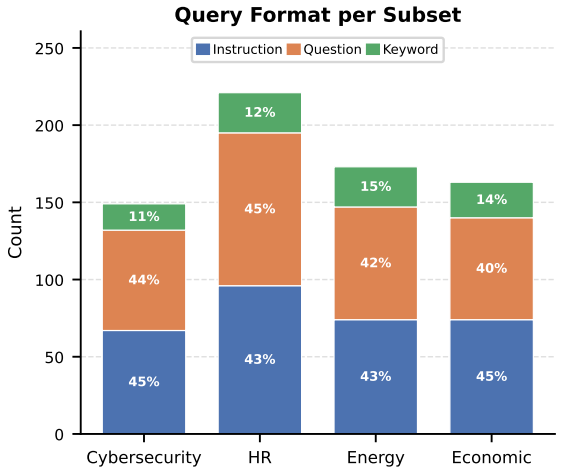


Figure 4: Distribution of query formats across subsets.

Subset	#Docs	#Pages	#Queries	#Qrels	Avg. Pages / Query
Cybersecurity	17	1,150	149	409	2.74
HR	9	2,109	221	726	3.30
Energy	11	1,993	173	525	3.03
Economic	20	1,477	163	413	2.55
Total	57	6,729	706	2,073	2.94

Table 1: Statistics of the KoViDoRe benchmark across domain-specific subsets. Avg. Pages / Query denotes the average number of relevant pages per query.

timodal and multi-page settings. In addition, we remove low-quality queries such as those with excessive keyword enumeration or those that directly reveal answer content from the document, resulting in a reduced but more reliable set of keyword-based queries. For the benchmark, we further incorporate human verification, where annotators perform a final review of query-page relevance annotations and refine queries through rephrasing when they are unnatural or ambiguous. This process balances automatic filtering and human verification to produce high-quality annotations.

3.5 Dataset Statistics

Table 1 summarizes the overall statistics of the benchmark across its four domain-specific subsets. Figure 3 and Figure 4 further illustrate the distributions of query types and query formats. The benchmark consists of 57 documents and 6,729 pages, with a total of 706 queries and 2,073 relevance annotations. The four domain-specific subsets exhibit notable differences in scale and structure. For example, the HR and Energy subsets contain a larger number of pages per query, suggesting that queries in these domains often require aggregating information from multiple pages. As shown in Figure 3 and Figure 4, the dataset contains a diverse set of query types and formats across all subsets. Multi-hop, open-ended, and comparison-based queries appear frequently, while question- and instruction-style queries are more common than keyword queries.

4 Experiments

4.1 Experimental Setup

We evaluate Korean visual document retrieval as a ranking task over document pages. Given a query q and a collection of document pages $\mathcal{D}$, the goal is to retrieve and rank pages that are relevant to the query. Each document is represented as a set of pages containing both textual and visual content. Queries are written in Korean and reflect diverse information needs grounded in real-world documents. Relevance is defined at the page level with graded labels: a page is labeled as fully relevant (2) if it contains sufficient information to answer the query, and partially relevant (1) if it provides supporting evidence. The dataset follows the BEIR-style evaluation framework with a corpus, queries, and relevance judgments (qrels) (Thakur et al., 2021). We report performance using nDCG@10. We evaluate a range of multimodal retrieval models on the KoViDoRe benchmark, covering small ($<$1B), medium (1B–4B), and large ($\geq$4B) models. The evaluated models include CLIP-based encoders (Radford et al., 2021; Zhai et al., 2023; Koukounas et al., 2024), late-interaction retrieval models such as ColPali and ColQwen (Faysse et al., 2024; Nomic Team, 2025; Huang and Tan, 2025), and recent multimodal embedding models (Jiang et al., 2024; Günther et al., 2025; Li et al., 2026). Evaluation is conducted across four domains: Cybersecurity, Energy, Economic, and Human Resources.

We conduct evaluation using the MTEB (Massive Text Embedding Benchmark) (Muennighoff et al., 2022), adapted to support multimodal retrieval on the KoViDoRe dataset. This provides a standardized and reproducible evaluation pipeline across all models. We build separate retrieval indices for each domain subset and each query is evaluated only against the pages within its corresponding domain.

4.2 Main Results

Table 2 presents the performance of all evaluated models on the benchmark. Overall, we observe a clear performance gap between model scales. Small models perform poorly across all domains, often failing to retrieve relevant pages. Medium-scale models show improved performance, particularly those based on late-interaction architectures. Larger models generally achieve higher scores, although performance gains are not uniform across domains. Among all models, `jina-embeddings-v4`

Model	Params	Cyber	Energy	Economic	HR	Avg.
Small Models (<1B parameters)						
openai/clip-vit-base-patch16♠	151M	4.1	0.8	0.0	0.6	1.4
vidore/colSmol-256M♦	256M	19.7	9.5	1.0	1.1	7.8
vidore/colSmol-500M♦	500M	26.2	9.9	0.6	0.9	9.4
jinaai/jina-clip-v2♠	865M	20.4	11.3	0.2	3.1	8.8
google/siglip-so400m-patch14-384♠	878M	15.3	5.3	1.3	1.1	5.8
Medium Models (1B–4B parameters)						
Qwen/Qwen3-VL-Embedding-2B♣	2.0B	61.3	40.6	15.3	18.7	34.0
vidore/colqwen2-v1.0♦	2.2B	53.3	42.0	8.0	14.7	29.5
vidore/colpali-v1.1♦	2.9B	31.9	18.2	3.0	6.0	14.8
vidore/colpali-v1.2♦	2.9B	33.2	16.4	2.1	4.5	14.1
vidore/colpali-v1.3♦	2.9B	34.7	20.6	1.6	6.2	15.8
ApsaraStackMaaS/EvoQwen2.5-VL-Retriever-3B-v1♦	3.0B	41.4	31.5	6.3	11.3	22.6
nomic-ai/colnomic-embed-multimodal-3b♦	3.0B	47.4	44.2	10.5	32.9	33.7
vidore/colqwen2.5-v0.2♦	3.0B	43.9	44.3	3.9	13.5	26.4
jinaai/jina-embeddings-v4♣	3.8B	<u>77.6</u>	**67.7**	**24.5**	**50.1**	**55.0**
Large Models (≥4B parameters)						
TomoroAI/tomoro-colqwen3-embed-4b♦	4.0B	55.3	31.0	9.1	10.1	26.4
eagerworks/eager-embed-v1♣	4.0B	51.5	32.7	5.4	7.0	24.2
TIGER-Lab/VLM2Vec-Full♣	4.2B	9.8	3.2	1.3	1.3	3.9
ApsaraStackMaaS/EvoQwen2.5-VL-Retriever-7B-v1♦	7.0B	66.0	55.4	12.1	26.4	40.0
nomic-ai/colnomic-embed-multimodal-7b♦	7.0B	69.6	59.5	12.4	33.3	43.7
Qwen/Qwen3-VL-Embedding-8B♣	8.0B	**77.8**	<u>63.2</u>	<u>23.4</u>	<u>37.4</u>	<u>50.4</u>
TomoroAI/tomoro-colqwen3-embed-8b♦	8.0B	73.7	58.5	16.3	26.5	43.8

Table 2: Performance comparison on KoViDoRe benchmark (nDCG@10, %). **Bold** indicates the best score; <u>underline</u> indicates the second-best score. ♠: CLIP-based, ♦: late-interaction, ♣: single-vector models. Cyber: Cybersecurity, HR: Human Resources.

achieves the best overall performance, significantly outperforming other models across all domains. This suggests that retrieval effectiveness is influenced not only by model scale, but also by factors such as training objective and data composition.

4.3 Analysis

Despite improvements from larger models, performance remains limited across all domains. In particular, domains such as Economic and Human Resources consistently show lower scores, indicating that retrieving relevant information in these settings is especially challenging. This is likely due to more complex document structures and the presence of information distributed across multiple pages.

We also observe that even strong retrieval models achieve relatively limited performance on several subsets of KoViDoRe. This suggests that the benchmark introduces additional challenges beyond conventional document retrieval settings, including complex layouts, structured visual content, and information distributed across multiple pages. In particular, queries associated with multiple relevant pages remain difficult for existing models, indicating that effectively retrieving and aggregating distributed document evidence is still a challenging problem in Korean visual document retrieval.

Motivated by this limitation, we further investigate whether training on Korean-specific data can improve retrieval performance, which we explore in the following subsection.

4.4 Ko-VDR Train Public

Dataset Construction To address the limitations identified in the previous section, we curate a large-scale training dataset, **Ko-VDR Train Public**, using the same data generation pipeline. The dataset consists of query-page pairs derived from Korean visual documents and includes a total of **310,226** query-page pairs. To ensure data quality, we apply both consistency-based and rule-based filtering. The consistency-based filtering retains only query-page pairs where relevance signals from the query generation stage and the additional relevance mapping stage agree. In addition, we apply rule-based filtering to remove low-quality queries, including those with excessive keyword enumeration and those that directly reveal answer content from

Model	Params	Cyber	Energy	Economic	HR	Avg.
vidore/colSmol-500M	500M	26.2	9.9	0.6	0.9	9.4
+ Ko-VDR Train Public		39.4	35.0	14.4	18.7	26.9
vidore/colqwen2-v1.0	2.2B	53.3	42.0	8.0	14.7	29.5
+ Ko-VDR Train Public		75.6	67.6	18.3	49.6	52.8
jinaai/jina-embeddings-v4	3.8B	77.6	**67.7**	**24.5**	**50.1**	**55.0**
TomoroAI/tomoro-colqwen3-embed-4b	4.0B	55.3	31.0	9.1	10.1	26.4
Qwen/Qwen3-VL-Embedding-8B	8.0B	**77.8**	63.2	23.4	37.4	50.4

Table 3: Comparison with representative retrieval baselines on KoViDoRe (nDCG@10, %). **Bold** indicates the best score; underline indicates the second-best score. Cyber: Cybersecurity, HR: Human Resources.

the document. This helps eliminate trivial or overly extractive cases and improves the robustness of the data. Unlike the benchmark construction process, we do not perform additional human verification for the training dataset to maintain scalability.

Training Setup We fine-tune two late-interaction retrieval models, `colSmol-500M` and `colqwen2-v1.0`, using the `colpali_engine` framework. Training is conducted on 2× NVIDIA B200 GPUs using BF16. We use a batch size of 128 per device and train for 3 epochs. In addition to Ko-VDR Train Public, we mix in a private Korean visual QA dataset containing TableVQA- and FigureVQA-style supervision (Kim et al., 2024; Kahou et al., 2017). This additional dataset complements the retrieval objective by providing stronger supervision for structured visual understanding, particularly for tables and figures.

Results Table 3 shows that fine-tuning on our dataset consistently improves performance across both `colSmol-500M` and `colqwen2-v1.0`. The smaller `colSmol-500M` model achieves substantial gains across all domains, demonstrating the effectiveness of the proposed training data even for lightweight models. For `colqwen2-v1.0`, fine-tuning leads to significant performance improvements, achieving competitive results with strong multimodal embedding models and surpassing `Qwen3-VL-Embedding-8B` despite being smaller in scale. These results highlight that training on Korean-specific data can substantially improve retrieval performance and enable smaller models to compete with larger counterparts.

4.5 Interpretability

To better understand model behavior, we visualize query-to-document similarity using heatmaps for the fine-tuned `colqwen2-v1.0` model. In Figure 5,

Query: 독일 현물시장 참가자 수 감소가 **선물시장** 비중 확대와 시장 구조 분화와 관련이 있나요?

Figure 5: Similarity map on a document example.

the model assigns high similarity to the term "선물시장," indicating that it correctly focuses on the key textual evidence relevant to the query. This suggests that the model is able to identify and attend to query-relevant terms in Korean visual documents. Additional examples are provided in Figure 15.

5 Ablation Study

5.1 Effect of the Number of Relevant Pages

To analyze how retrieval performance varies with query complexity, we group queries by the number of relevant pages and report nDCG@10 for each group. Figure 6 shows the results for `Qwen3-VL-Embedding-2B` and `Qwen3-VL-Embedding-8B`. We observe a general trend where performance tends to decrease as the number of relevant pages increases. For both models, performance is highest when a query is

Model	Cyber	Energy	Economic	HR	Avg.
vidore/colqwen2-v1.0	53.3	42.0	8.0	14.7	29.5
+ Private Only	70.3	58.8	**18.7**	37.7	46.4
+ Public Only	<u>75.4</u>	<u>66.9</u>	16.5	<u>49.3</u>	<u>52.0</u>
+ **Private + Public**	**75.6**	**67.6**	<u>18.3</u>	**49.6**	**52.8**

Table 4: Effect of training data composition on `vidore/colqwen2-v1.0` (nDCG@10, %). **Bold** indicates the best score; <u>underline</u> indicates the second-best score. Cyber: Cybersecurity, HR: Human Resources.

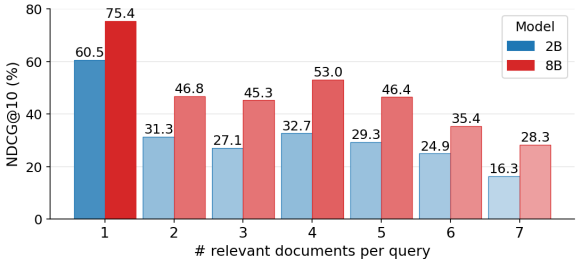


Figure 6: nDCG@10 by the number of relevant pages per query. Performance tends to decrease as the number of relevant pages increases.

associated with a single relevant page, and generally declines as more relevant pages are required, although the decrease is not strictly monotonic. This suggests that queries requiring evidence from multiple pages are more challenging, as models must retrieve and integrate information distributed across different parts of a document. While `Qwen3-VL-Embedding-8B` consistently outperforms `Qwen3-VL-Embedding-2B` across all groups, both exhibit similar trends, indicating that increasing model capacity alone does not fully mitigate the challenges of multi-page retrieval. These results highlight that KoViDoRe captures the increased difficulty of queries with broader information needs, providing a more realistic evaluation setting in which retrieval systems must aggregate evidence across multiple pages.

5.2 Effect of Training Data Composition

We analyze the effect of training data composition using `vidore/colqwen2-v1.0`. Following the training setup described in Section 4.4, we keep all training configurations fixed and vary only the composition of the training data. As shown in Table 4, training with private data alone improves performance over the base model, indicating the benefit of Korean-specific supervision. Training with public data yields larger improvements across most subsets, suggesting that broader data coverage and diversity contribute significantly to retrieval performance on Korean visual documents. When both private and public data are combined, the model achieves the best overall performance across most subsets, demonstrating that Korean-specific supervision and large-scale public training data are complementary. In particular, combining both datasets consistently improves performance in Cybersecurity, Energy, and HR, leading to the highest average performance overall. Interestingly, the Economic subset exhibits a different trend, where training with private data alone achieves the highest performance. We hypothesize that this is because many queries in the Economic subset require identifying and interpreting complex multi-column tables distributed across document pages. Since the private dataset includes TableVQA-style supervision, it likely provides stronger training signals for structured table understanding, resulting in larger gains on table-heavy economic documents.

6 Conclusion

We introduced KoViDoRe, a benchmark for Korean visual document retrieval. Unlike conventional benchmarks that primarily focus on queries answerable from a single page, KoViDoRe emphasizes queries that require retrieving and integrating information distributed across multiple pages. We constructed the dataset from publicly available Korean documents with diverse layouts and developed a LLM-based multi-stage pipeline with human verification. Through extensive evaluation, we showed that current multimodal retrieval models struggle to effectively handle Korean visual document retrieval, particularly in scenarios involving structured content and diverse query types. To address this limitation, we further curated Ko-VDR Train Public, a large-scale training dataset designed for Korean visual document retrieval. Our experiments demonstrate that training on Korean-specific data improves retrieval performance, highlighting the importance of language-specific training resources. We hope that KoViDoRe and Ko-VDR Train Public will facilitate future research on Korean visual documents retrieval.

Limitations

Despite the contributions of this work, several limitations remain. First, our query generation relies on parsed markdown representations and image captions rather than raw visual inputs. While this design enables scalable and reproducible data generation, it may not fully preserve fine-grained visual information such as layout, color, or chart-specific patterns. As a result, information loss may occur in visually intensive documents, potentially affecting query quality. As future work, we plan to incorporate vision-language models (VLMs) into the query generation process to better capture visual information and reduce such information loss. Second, while our relevance mapping process reduces manual annotation effort through consistency-based filtering, it may still introduce noise due to imperfect alignment between generation and mapping stages. In addition, although the private Korean VQA dataset used in our training experiments contributes to performance improvements, it cannot be publicly released due to licensing restrictions. As a result, the fully reproducible training setup is limited to the publicly available components. Finally, our experiments focus on evaluating existing retrieval models, and we do not propose new model architectures specifically designed for Korean visual document retrieval. Future work may explore model designs and training strategies better suited for handling structured and visually rich documents, including approaches that directly incorporate visual inputs during query generation.

Acknowledgments

This work was supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT) (IITP-2026-RS-2024-00438239).

References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16776–16809, Vienna, Austria. Association for Computational Linguistics.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multimodal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*.

Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. 2025. MMDocIR: Benchmarking multimodal retrieval for long documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30971–31005, Suzhou, China. Association for Computational Linguistics.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and 1 others. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550.

Xin Huang and Kye Min Tan. 2025. Beyond text: Unlocking true multimodal, end-to-end rag with tomoro colqwen3.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2024. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*.

Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. 2024. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*.

Jaehoon Lee, Sohyun Kim, Wanggeun Park, Geon Lee, Seungkyung Kim, and Minyoung Lee. 2025. Sds kopub vdr: A benchmark dataset for visual document retrieval in korean public documents. *arXiv preprint arXiv:2511.04910*.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Chen Keqin, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun

Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*.

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. 2026. Vidore v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. *arXiv preprint arXiv:2601.08620*.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, Miami, Florida, USA. Association for Computational Linguistics.

Quentin Macé, António Loison, and Manuel Faysse. 2025. Vidore benchmark v2: Raising the bar for visual retrieval. *arXiv preprint arXiv:2505.17166*.

Gabriel de Souza P Moreira, Ronay Ak, Mengyao Xu, Oliver Holworthy, Benedikt Schifferer, Zhiding Yu, Yauhen Babakhin, Radek Osmulski, Jiarui Cai, Ryan Chesler, and 1 others. 2026. Nemotron colembed v2: Top-performing late interaction embedding models for visual document retrieval. *arXiv preprint arXiv:2602.03992*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Nomic Team. 2025. Nomic embed multimodal: Interleaved text, image, and screenshots for visual document retrieval.

Radek Osmulski, Gabriel de Souza P Moreira, Ronay Ak, Mengyao Xu, Benedikt Schifferer, and Even Oldridge. 2025. Miracl-vision: A large, multilingual, visual document retrieval benchmark. *arXiv preprint arXiv:2505.11651*.

Xiangyu Peng, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, and Chien-Sheng Wu. 2025. Unidocbench: A unified benchmark for document-centric multimodal rag. *arXiv preprint arXiv:2510.03663*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Connor Shorten, Augustas Skaburskas, Daniel M Jones, Charles Pierse, Roberto Esposito, John Trengrove, Etienne Dilocker, and Bob van Luijt. 2026. Irpapers: A visual document benchmark for scientific retrieval and question answering. *arXiv preprint arXiv:2602.17687*.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. 2025. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5311–5329.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9124–9145.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. 2025. REAL-MM-RAG: A real-world multi-modal retrieval benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31660–31683, Vienna, Austria. Association for Computational Linguistics.

Zilin Xiao, Qi Ma, Mengting Gu, Chun-cheng Jason Chen, Xintao Chen, Vicente Ordonez, and Vijai Mohan. 2025. Metaembed: Scaling multimodal retrieval at test-time with flexible late interaction. *arXiv preprint arXiv:2509.18095*.

Yibo Yan, Jiahao Huo, Guanbo Feng, Mingdong Ou, Yi Cao, Xin Zou, Shuliang Liu, Yuanhuiyi Lyu, Yu Huang, Jungang Li, and 1 others. 2026. Unlocking multimodal document intelligence: From current triumphs to future frontiers of visual document retrieval. *arXiv preprint arXiv:2602.19961*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

A Appendix

This appendix provides supplementary materials for transparency and reproducibility. We include a detailed comparison with the SDS KoPub VDR benchmark, the technical rationale for our model selections, and formal category definitions used in query generation. Additionally, we provide the complete set of prompt templates for key pipeline stages, representative query-page examples, and comprehensive document metadata for all source collections included in KoViDoRe.

A.1 Comparison to SDS KoPub VDR

While SDS KoPub VDR (Lee et al., 2025) represents an important benchmark for evaluating Korean visual document retrieval, KoViDoRe introduces a different task formulation that emphasizes more complex retrieval scenarios. The primary distinction lies in the query-to-document mapping: SDS KoPub VDR is largely designed for single-page retrieval, where each query is mapped to a single relevant page. In contrast, KoViDoRe explicitly focuses on multi-page evidence aggregation, with each query associated with an average of 2.94 relevant pages. This shift from single-page matching to multi-page evidence gathering aligns with realistic enterprise search scenarios, where information is often distributed across multiple pages and no single page alone is sufficient to satisfy the query. As shown in Table 1, KoViDoRe's queries frequently require synthesizing information from diverse document regions, such as comparing financial trends or summarizing policies spread throughout a report. By providing a higher ratio of relevant pages per query—reaching up to 3.30 in the HR subset—KoViDoRe complements existing resources by evaluating a model's ability to retrieve information from multiple pages to satisfy the diverse informational needs embedded in a single query.

A.2 Document Parsing and Query Generation Models

Upstage Document Parse We use Upstage Document Parse as the document parsing backend in our pipeline. The model provides layout-aware parsing of visually rich documents, extracting both page-level and element-level markdown representations, and decomposing each page into structured components such as text blocks, tables, figures, charts, and diagrams. It also generates captions for visual elements, enabling semantic interpretation of non-textual content. According to publicly reported results on DP-Bench, the model achieves strong performance in preserving document structure, and is designed to handle complex document layouts. We choose this model as it is effective for parsing structurally complex Korean documents while preserving layout-aware information, which is critical for downstream query generation and relevance mapping.

Solar-Pro3 For query generation, we use Solar-Pro3 as the underlying language model. Solar-Pro3 is a reasoning-oriented language model designed to support structured and context-aware generation. According to publicly available reports, it demonstrates strong performance in Korean language understanding and instruction-following tasks. Given document content in markdown format, the model generates queries conditioned on both textual and visual information extracted during preprocessing. We choose Solar-Pro3 as it is well-suited for generating complex Korean queries that require multi-step reasoning, which aligns with the objective of constructing realistic and challenging retrieval scenarios in KoViDoRe.

A.3 Query Category Definitions

Table 5 and Table 6 define the query type and query format categories used in our dataset construction pipeline.

A.4 Prompt Templates

Figures 9, 10, 11, and 12 present the prompt templates used for summary generation, relevance mapping, and query generation.

A.5 Example Query-Page Pairs

Figure 13 and Figure 14 show representative examples of query-page pairs from different subsets of KoViDoRe.

A.6 Document Metadata

Table 7 lists the metadata of the document collections used in KoViDoRe, while Table 8 presents the metadata of the documents used in Ko-VDR Train Public. Both tables include document titles, providers, page counts, and license information.

A.7 Distribution of Relevant Pages per Query

Figure 7 shows the distribution of queries with respect to the number of annotated relevant pages.

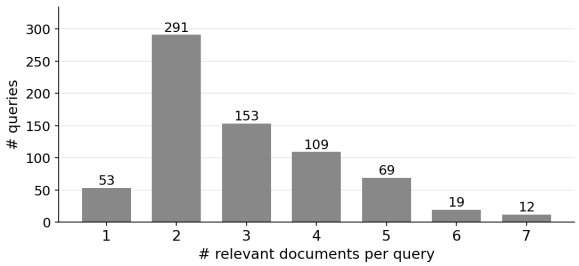


Figure 7: Distribution of queries by the number of relevant pages.

Most queries are associated with two or three relevant pages, while queries requiring a larger number of relevant pages are less frequent. This distribution reflects a realistic setting where information needs vary in complexity, with a substantial portion of queries requiring aggregation across multiple pages. At the same time, the presence of queries with a higher number of relevant pages supports the analysis in Section 5.1, demonstrating that KoViDoRe includes challenging cases that require broader evidence aggregation.

Category	Definition
open-ended	A query requiring synthesis and explanation of information. The answer must integrate multiple concepts into a coherent narrative rather than citing a single fact.
compare-contrast	A query requiring identification and articulation of similarities and/or differences between two or more entities, concepts, or topics.
enumerative	A query requesting a complete list of items that meet specific criteria.
numerical	A query expecting a numerical value, obtained either by direct extraction or calculation.
boolean	A query expecting a yes/no answer, potentially requiring reasoning over extracted information.
extractive	A query answerable by directly citing a specific fact or piece of information from the documents.
multi-hop	A query requiring information retrieval from multiple distinct sources or sections, which must then be combined to produce a complete answer.

Table 5: Query type categories and their definitions.

Category	Definition
question	A query presented in the form of a direct question, seeking specific information or clarification.
instruction	A query framed as a directive or command, requesting the model to perform a specific task or provide information in a particular manner.
keyword	A query consisting of noun phrases and keywords only, WITHOUT forming a complete sentence. No verbs, no question words, no sentence endings. Mimics how users type into search engines: fragmented, concise, noun-centric terms separated by spaces.

Table 6: Query format categories and their definitions.

```
You are a document analysis expert. Your role is to analyze a specific section of a document
within the context of the entire page and summarize its core content in Korean.

## Context and Task:
- You are provided with the full page content in markdown format.
- You are also provided with a specific section's data, which may include text segments,
coordinates, and metadata.
- Your task is to summarize the core information of the **Target Section ** while using the
**Full Page** as contextual background to ensure accuracy and completeness.

## Summary Principles:
- Capture all key information, concepts, and data specifically related to the Target Section.
- If the Target Section refers to tables, charts, or graphs present in the Full Page,
describe their content and significance.
- Ensure that numerical data, statistics, and important figures from the section are
strictly included.
- The summary should be 5-7 sentences long—concise, yet minimizing any loss of information.
- A reader should be able to understand the core message of the specific section solely by
reading the summary.

## Output Format:
- Output only the summary written in Korean.
- Do not provide any introductory remarks, preambles, or additional explanations.

## Instructions:
Please summarize the Target Section based on the Full Page context.

### Full Page Markdown:
{{ markdown }}

### Target Section:
{{ elements }}
```

Figure 8: Prompt for generating single-section summaries based on full-page context

```
You are a document synthesis and integration expert. Your role is to analyze a set of
fragmented summaries extracted from different pages of a single document and synthesize them
into a coherent, comprehensive cross-section summary in Korean.

## Context and Task:
- You are provided with a **Combined Context**, which consists of multiple summaries derived
from randomly sampled sections of a document.
- The context is listed with page numbers to indicate where each information originates.
- Your task is to generate a **Cross-Section Summary** that integrates these dispersed
pieces of information into a unified narrative.
- You must identify logical connections, thematic consistency, or causal relationships
between the sections, even if the page numbers are not consecutive.

## Summary Principles:
- **Integration over Listing:** Do not simply list the summaries one by one. Instead, weave
them together to explain "what this document is discussing" based on the available evidence.
- **Contextual Flow:** Use the page numbers to infer the structure (e.g., "The document
introduces [Topic] on Page 2 and later elaborates on [Specific Detail] on Page 15").
- **Handling Gaps:** Acknowledge that the information is sampled. If sections seem unrelated,
describe them as distinct aspects covered within the document.
- **Accuracy:** Strictly adhere to the provided content. Do not hallucinate information not
present in the input snippets.
- **Length & Tone:** Write a professional, dense paragraph (7-10 sentences). Use a formal
and objective tone.

## Output Format:
- Output only the summary written in Korean.
- Do not provide any introductory remarks, preambles, or additional explanations.

## Instructions:
Please summarize the core message based on the provided Combined Context below.

### Combined Context:
<sections>
{% for summary in single_section_summary %}
<section index="{{ loop.index0 }}">
{{ summary }}
</section>
{% endfor %}
</sections>
```

Figure 9: Prompt for cross-section summarization using aggregated section summaries

```
You are a strategic Document Relevance Auditor. Your goal is to identify pages that provide
either a "Complete Answer" or "Essential Building Blocks" for a query.

## Task
Evaluate the relevance of each document page. You must distinguish between "Noisy/Empty
pages" and "Partial but Crucial data pages."

## Query
{{ query }}

## Documents
<documents>
{% for doc in markdown %}
<document index="{{ loop.index0 }}">
{{ doc }}
</document>
{% endfor %}
</documents>

## Scoring Criteria (Balanced Evidence-Based)

- **2 (FULLY_RELEVANT)**: The page contains an explicit, direct, and complete answer to all
parts of the query.
- **1 (CRITICALLY_RELEVANT)**: The page contains specific, substantive facts or data
required to answer *at least one part* of a multi-part query.
    - (e.g., If the query asks for "A and B comparison" and the page has detailed data on
    "A", it is a CRITICAL building block, even if "B" is missing.)
    - (e.g., Detailed statistics, specific policy names, or factual descriptions that would
    form part of the final answer.)
- **0 (IRRELEVANT)**: The page provides no substantive value. This includes:
    - **Pure Navigation**: Tables of Contents or cover pages with only titles/page numbers.
    - **Off-Topic**: Content that doesn't address any specific component of the query.
    - **Vague Mentions**: Just mentioning a keyword without any descriptive facts or data.

## Critical Instructions
1. **The Building Block Rule**: Do not reject a page just because it is incomplete. If it
provides a "Hard Fact" (e.g., China's specific Metaverse policy) that is part of the query's
scope, assign 1.
2. **Substance Over Format**: A table or a list of policies is highly relevant if it contains
the "What/How/When" of the subject, even if it doesn't "compare" it for you.
3. **Anti-Hallucination**: While being more inclusive of partial data, still score 0 if the
page requires you to "guess" the information. The data must be explicitly written.

## Reasoning Requirements (Thinking Process)
For each page, explain your judgment in **KOREAN**:
- **Partial match check**: Does this page cover at least one specific component of the query?
- **Fact density**: Does it provide concrete data/facts, or just general mentions?
- **Role in Answer**: How does this information help in constructing the final response?

## Output Format
Return your assessment with:
1. `reasoning`: A detailed **KOREAN** explanation for each page.
2. `relevance_scores`: A list of integer scores (0, 1, or 2).
```

Figure 10: Prompt for relevance mapping by identifying fully relevant and critically relevant pages

```
You are an expert in creating challenging datasets for Vision Document Retrieval (VDR).
Your goal is to generate a **highly specific Korean search query** that acts as a realistic
user prompt for retrieving information from a large corpus.

### 1. Document Context
You are provided with two types of summaries:

#### 1.1 Single-Section Summaries
Each <single_section_summary> tag contains a summary of a **specific section/page**,
identified by its **actual page number** (page).

<single_section_summaries>
{% for summary in single_section_summary %}
<single_section_summary index="{{ loop.index0 }}">
{{ summary }}
</single_section_summary>
{% endfor %}
</single_section_summaries>

#### 1.2 Cross-Section Summary
The following is a **synthesized summary** that integrates information across all the
sections above. Use this to understand the overall narrative and relationships between
different sections.

<cross_section_summary>
{{ cross_section_summary }}
</cross_section_summary>

**How to use these summaries**:
* Use **single-section summaries** to identify specific facts, entities, and details located
on each page.
* Use **cross-section summary** to understand how information connects across pages and to
identify synthesis opportunities.
* Your query should require combining specific details from multiple sections (identified via
single-section summaries) in a way that reflects the cross-section relationships (identified
via cross-section summary).

### 2. Task Requirements
You must generate a structured output containing the rationale and the query itself based on
the following specifications:

* **Query Type**: {{ query_type }} ({{ query_type_definition }})
* **Query Format**: {{ query_format }} ({{ query_format_definition }})

### 3. Critical Constraints for Realistic Retrieval

#### Rule 1: NO Artificial Location References
* **Strictly Forbidden**: "2페이지에서...", "다음 장에 있는...", "첫 번째 문서의...", "위에서
언급된..."
* **Reason**: The user queries the entire database and does not know the document order or
page numbers.
* **Alternative**: Use **Section Headers, Table Captions, or Unique Keywords** found in the
text.
  * Bad: "2페이지에 있는 표를 요약해."
  * Good: "'2024년 재무 하이라이트' 표를 요약해."

#### Rule 2: Implicit Multi-Page Synthesis
* The query must require information scattered across multiple pages, but **without
explicitly stating so**.
* **Strategy**: Identify **Entity A** on one page and **Entity B** on another, then ask about
their relationship.

#### Rule 3: Entity-Grounded Specificity
* Avoid generic queries like "에너지 정책을 분석해줘."
* Include specific entities found in the text: **Dates, Company Names, Regulations (e.g.,
ISO-27001), Project Codes, Policy Names, or Program Names**.
* However, do NOT include exact numerical values (see Rule 5).

#### Rule 4: Single Natural Query
* The query **MUST be a single unit** appropriate to its format (one question, one
instruction, or one keyword cluster).
* **Strictly Forbidden Patterns**:
  * Multiple sentences: "~입니다. ~해주세요."
```

  * Instruction suffixes: "단, ~를 기준으로 답변하시오."
  * Explicit output format requests: "~를 근거로 제시하시오.", "~를 나열하시오."
  * Conditional clauses at the end: "단, ~를 구분하여 제공해야 합니다."
* **Examples**:
  * Bad: "2021년과 2022년 상승률을 비교하시오. 단, 수도권과 지방을 구분하여 제시하시오."
  * Good: "2021년과 2022년 수도권 및 지방의 주택 매매가격 상승률은 어떻게 달랐나요?"

Rule 5: Realistic Search Behavior
The query must read as if a **researcher who does NOT have the document** is searching a database by topic and keywords. This single rule covers three aspects:

(a) No Verbatim Document Data
* Use **conceptual references** (policy names, years, entity names) instead of **exact figures**.
* **Strictly Forbidden**: Specific monetary values, exact percentages, precise statistics copied from the document.
  * Bad: "에너지 요금이 €49.5/MWh에서 €94/MWh로 89% 상승한 이유는?"
  * Good: "2022년 프랑스 소매 에너지 요금 급등과 EDF의 ARENH 정책은 어떤 관계가 있나요?"

(b) No Document-Aware Framing
* **Strictly Forbidden**: "문서에서", "해당 자료의", "위 표에 따르면", "본 보고서의", "제시된 데이터를 기반으로"
* Also forbidden — **Document Title Scoping** (assumes the user already knows the document exists):
  * Bad: "제7차 에너지기본계획에서 원전 비중 목표는?"
  * Good: "2025년 일본의 원전 비중 목표"

(c) Realistic User Knowledge
* The user **knows**: topic area, key entities, time periods of interest.
* The user **does NOT know**: page numbers, document structure, specific numerical values, exact document titles.

4. Query Format Specification

Question Format (질문형)
* Must be a complete interrogative sentence with question endings.
* **Required elements**: Question word (무엇, 어떻게, 왜, 어떤) OR question ending (~인가요?, ~있나요?, ~했는가?)
* **Examples**:
  * "M2 광의통화 증가율이 2020년 국가채무 증가에 영향을 미쳤는가?"
  * "에너지바우처 제도의 지원 대상은 누구인가요?"

Instruction Format (지시형)
* Must be a command with imperative endings.
* **Required elements**: Imperative ending (~해주세요, ~하시오, ~분석하라, ~설명하라)
* **Examples**:
  * "2020년 M2 통화량과 국가채무 간의 상관관계를 분석해주세요."
  * "에너지바우처와 에너지효율개선 사업의 차이점을 비교하라."

Keyword Format (키워드형)
* **NO complete sentences.** Only noun phrases and search terms.
* **NO verbs, NO question words, NO sentence endings.**
* Mimics search engine input: fragmented, noun-centric.

Keyword Format Rules:
Allowed / Forbidden
- 명사, 명사구, 복합 명사구 / 동사 (~하다, ~이다, ~있다)
- 관계 조사 (~의, ~와/과, ~간, ~에 따른, ~으로 인한) / 질문사 (무엇, 어떻게, 왜, 어떤)
- 고유명사, 연도, 날짜 / 문장 종결 (~인가요, ~해주세요, ~입니까)
- 관계 표현 (비교, 관계, 영향, 연관성, 상관관계) / 완전한 문장 구조
- 영문 약어 (EDF, ARENH, GDP) / 공백으로만 나열된 독립 키워드들
- 개념적 추상화 표현 / 문서 표 항목명·인덱스의 직접 복사

Keyword Structural Templates
A keyword query must form **one coherent noun phrase**. Every noun must be connected to its neighbors by Korean particles (의, 와/과, 간, 에 따른, 으로 인한, 내, 중, 및) that make the semantic relationship explicit.

Templates:
- Comparison: A의 X와/과 B의 Y (간) 차이/비교
  Example: "운수업의 부가가치당 에너지소비량과 수송용 에너지소비 비중 차이"
- Correlation: A와/과 B 간 연관성/관계/상관관계
  Example: "일반가구의 설계가중치와 도시가구의 에너지소비 간 연관성"

```
- Causation: A 변화/증가/감소와 B 변화의 연관성/영향
  Example: "부산 개별여행 비중 증가와 농수산물 구매 비중 상승의 연관성"
- Condition: A에 따른/으로 인한 B의 변화/추이
  Example: "스페인 용량요금 중단에 따른 전력부문 적자의 변화"
- Composition: A 내 B와 C의 비중/분포/구성
  Example: "EU 노동 인력 내 녹색 직업과 고도 디지털 집약 직업 간의 연령 분포"

**Particle Removal Test**:
Strip all particles (의/와/과/간/에 따른/으로 인한/내/중/및) from the query.
* If the meaning **collapses** -> Well-formed noun phrase.
* If the meaning **stays the same** -> Keyword bag. Rewrite.

**Read-Aloud Test**:
Read the query aloud. If there is a natural pause splitting it into two independent chunks
with no grammatical bridge -> Two queries glued together. Rewrite.

**Bad -> Fixed Examples**:
* "감일도서관 개관 희망도서 바로대출 지역서점 연계 독서문화 활성화 지원 사업 이동도서관
스마트도서관"
  -> "감일도서관 개관 이후 희망도서 바로대출 서비스와 지역서점 연계 독서문화 사업 간의 운영 방식
차이"
* "K-방산 폴란드 수출 비중 라틴아메리카 방위비 증가"
  -> "K-방산의 폴란드 수출 비중 확대와 라틴아메리카 방위비 증가 간 연관성"
* "베트남 최종 법인세 신고 베트남 개인소득세 체계 동일 과세 기준 여부"
  -> "베트남 법인세 최종 신고 체계와 개인소득세 체계의 과세 기준 동일 여부"

### 5. Quality Checklist (Self-Verification)
Before finalizing, verify ALL checks pass:
- Format Compliance: Query strictly follows the specified format
(question/instruction/keyword)
- Single Unit: ONE question, ONE instruction, or ONE keyword phrase — no multiple sentences
- No Page References: No page numbers, document indices, or positional references
- Realistic Search: No exact values from the document, no document-aware framing, no document
title scoping (Rule 5)
- Entity-Grounded: Includes searchable entities (names, years, policy names) but not verbatim
data
- Multi-Page Implicit: Requires information from multiple pages without explicitly stating it
- Keyword Coherence (keyword only): **Particle Removal Test** passes: stripping particles
must break the meaning
- Single Phrase (keyword only): **Read-Aloud Test** passes: query flows as one utterance with
no independent chunks

### 6. Output Generation
Generate the output strictly adhering to the defined JSON schema.
The query must be in **Korean** and must pass all checks in the Quality Checklist above.
**Pay special attention to the Query Format specification—the linguistic structure must
match exactly.**
```

Figure 11: Prompt for generating summary-based retrieval queries with controls for realism, diversity, and query formulation

```
You are an expert in creating challenging datasets for Vision Document Retrieval (VDR).
Your goal is to generate a **highly specific Korean search query** that acts as a realistic
user prompt for retrieving information from a large corpus.

### 1. Document Context
The following XML-like tags contain the markdown text extracted from a sequence of document
pages.
The <document index="..."> tags are for your internal reasoning ONLY. **Do NOT mention these
indices in the final query.**

<documents>
{% for doc in markdown %}
<document index="{{ loop.index0 }}">
{{ doc }}
</document>
{% endfor %}
</documents>

### 2. Task Requirements
You must generate a structured output containing the rationale and the query itself based on
the following specifications:

* **Query Type**: {{ query_type }} ({{ query_type_definition }})
* **Query Format**: {{ query_format }} ({{ query_format_definition }})

### 3. Critical Constraints for Realistic Retrieval

#### Rule 1: NO Artificial Location References
* **Strictly Forbidden**: "2페이지에서...", "다음 장에 있는...", "첫 번째 문서의...", "위에서
언급된..."
* **Reason**: The user queries the entire database and does not know the document order or
page numbers.
* **Alternative**: Use **Section Headers, Table Captions, or Unique Keywords** found in the
text.
  * Bad: "2페이지에 있는 표를 요약해."
  * Good: "'2024년 재무 하이라이트' 표를 요약해."

#### Rule 2: Implicit Multi-Page Synthesis
* The query must require information scattered across multiple pages, but **without
explicitly stating so**.
* **Strategy**: Identify **Entity A** on one page and **Entity B** on another, then ask about
their relationship.

#### Rule 3: Entity-Grounded Specificity
* Avoid generic queries like "에너지 정책을 분석해줘."
* Include specific entities found in the text: **Dates, Company Names, Regulations (e.g.,
ISO-27001), Project Codes, Policy Names, or Program Names**.
* However, do NOT include exact numerical values (see Rule 5).

#### Rule 4: Single Natural Query
* The query **MUST be a single unit** appropriate to its format (one question, one
instruction, or one keyword cluster).
* **Strictly Forbidden Patterns**:
  * Multiple sentences: "~입니다. ~해주세요."
  * Instruction suffixes: "단, ~를 기준으로 답변하시오."
  * Explicit output format requests: "~를 근거로 제시하시오.", "~를 나열하시오."
  * Conditional clauses at the end: "단, ~를 구분하여 제공해야 합니다."
* **Examples**:
  * Bad: "2021년과 2022년 상승률을 비교하시오. 단, 수도권과 지방을 구분하여 제시하시오."
  * Good: "2021년과 2022년 수도권 및 지방의 주택 매매가격 상승률은 어떻게 달랐나요?"

#### Rule 5: Realistic Search Behavior
The query must read as if a **researcher who does NOT have the document** is searching a
database by topic and keywords. This single rule covers three aspects:

**(a) No Verbatim Document Data**
* Use **conceptual references** (policy names, years, entity names) instead of **exact
figures**.
* **Strictly Forbidden**: Specific monetary values, exact percentages, precise statistics
copied from the document.
  * Bad: "에너지 요금이 €49.5/MWh에서 €94/MWh로 89% 상승한 이유는?"
  * Good: "2022년 프랑스 소매 에너지 요금 급등과 EDF의 ARENH 정책은 어떤 관계가 있나요?"
```

(b) No Document-Aware Framing
* **Strictly Forbidden**: "문서에서", "해당 자료의", "위 표에 따르면", "본 보고서의", "제시된 데이터를 기반으로"
* Also forbidden — **Document Title Scoping** (assumes the user already knows the document exists):
  * Bad: "제7차 에너지기본계획에서 원전 비중 목표는?"
  * Good: "2025년 일본의 원전 비중 목표"

(c) Realistic User Knowledge
* The user **knows**: topic area, key entities, time periods of interest.
* The user **does NOT know**: page numbers, document structure, specific numerical values, exact document titles.

4. Query Format Specification

Question Format (질문형)
* Must be a complete interrogative sentence with question endings.
* **Required elements**: Question word (무엇, 어떻게, 왜, 어떤) OR question ending (~인가요?, ~있나요?, ~했는가?)
* **Examples**:
  * "M2 광의통화 증가율이 2020년 국가채무 증가에 영향을 미쳤는가?"
  * "에너지바우처 제도의 지원 대상은 누구인가요?"

Instruction Format (지시형)
* Must be a command with imperative endings.
* **Required elements**: Imperative ending (~해주세요, ~하시오, ~분석하라, ~설명하라)
* **Examples**:
  * "2020년 M2 통화량과 국가채무 간의 상관관계를 분석해주세요."
  * "에너지바우처와 에너지효율개선 사업의 차이점을 비교하라."

Keyword Format (키워드형)
* **NO complete sentences.** Only noun phrases and search terms.
* **NO verbs, NO question words, NO sentence endings.**
* Mimics search engine input: fragmented, noun-centric.

Keyword Format Rules:
Allowed / Forbidden
- 명사, 명사구, 복합 명사구 / 동사 (~하다, ~이다, ~있다)
- 관계 조사 (~의, ~와/과, ~간, ~에 따른, ~으로 인한) / 질문사 (무엇, 어떻게, 왜, 어떤)
- 고유명사, 연도, 날짜 / 문장 종결 (~인가요, ~해주세요, ~입니까)
- 관계 표현 (비교, 관계, 영향, 연관성, 상관관계) / 완전한 문장 구조
- 영문 약어 (EDF, ARENH, GDP) / 공백으로만 나열된 독립 키워드들
- 개념적 추상화 표현 / 문서 표 항목명·인덱스의 직접 복사

Keyword Structural Templates
A keyword query must form **one coherent noun phrase**. Every noun must be connected to its neighbors by Korean particles (의, 와/과, 간, 에 따른, 으로 인한, 내, 중, 및) that make the semantic relationship explicit.

Templates:
- Comparison: A의 X와/과 B의 Y (간) 차이/비교
  Example: "운수업의 부가가치당 에너지소비량과 수송용 에너지소비 비중 차이"
- Correlation: A와/과 B 간 연관성/관계/상관관계
  Example: "일반가구의 설계가중치와 도시가구의 에너지소비 간 연관성"
- Causation: A 변화/증가/감소와 B 변화의 연관성/영향
  Example: "부산 개별여행 비중 증가와 농수산물 구매 비중 상승의 연관성"
- Condition: A에 따른/으로 인한 B의 변화/추이
  Example: "스페인 용량요금 중단에 따른 전력부문 적자의 변화"
- Composition: A 내 B와 C의 비중/분포/구성
  Example: "EU 노동 인력 내 녹색 직업과 고도 디지털 집약 직업 간의 연령 분포"

Particle Removal Test:
Strip all particles (의/와/과/간/에 따른/으로 인한/내/중/및) from the query.
* If the meaning **collapses** -> Well-formed noun phrase.
* If the meaning **stays the same** -> Keyword bag. Rewrite.

Read-Aloud Test:
Read the query aloud. If there is a natural pause splitting it into two independent chunks with no grammatical bridge -> Two queries glued together. Rewrite.

Bad -> Fixed Examples:
* "감일도서관 개관 희망도서 바로대출 지역서점 연계 독서문화 활성화 지원 사업 이동도서관 스마트도서관"

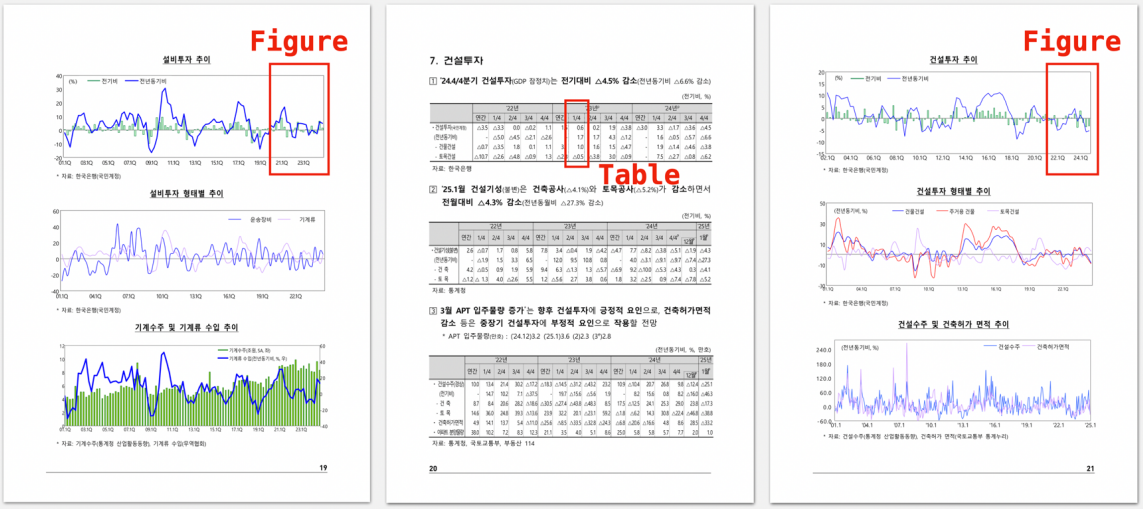
```
  -> "감일도서관 개관 이후 희망도서 바로대출 서비스와 지역서점 연계 독서문화 사업 간의 운영 방식
차이"
* "K-방산 폴란드 수출 비중 라틴아메리카 방위비 증가"
  -> "K-방산의 폴란드 수출 비중 확대와 라틴아메리카 방위비 증가 간 연관성"
* "베트남 최종 법인세 신고 베트남 개인소득세 체계 동일 과세 기준 여부"
  -> "베트남 법인세 최종 신고 체계와 개인소득세 체계의 과세 기준 동일 여부"

### 5. Quality Checklist (Self-Verification)
Before finalizing, verify ALL checks pass:
- Format Compliance: Query strictly follows the specified format
(question/instruction/keyword)
- Single Unit: ONE question, ONE instruction, or ONE keyword phrase - no multiple sentences
- No Page References: No page numbers, document indices, or positional references
- Realistic Search: No exact values from the document, no document-aware framing, no document
title scoping (Rule 5)
- Entity-Grounded: Includes searchable entities (names, years, policy names) but not verbatim
data
- Multi-Page Implicit: Requires information from multiple pages without explicitly stating it
- Keyword Coherence (keyword only): **Particle Removal Test** passes: stripping particles
must break the meaning
- Single Phrase (keyword only): **Read-Aloud Test** passes: query flows as one utterance with
no independent chunks

### 6. Output Generation
Generate the output strictly adhering to the defined JSON schema.
The query must be in **Korean** and must pass all checks in the Quality Checklist above.
**Pay special attention to the Query Format specification—the linguistic structure must
match exactly.**
```

Figure 12: Prompt for generating context-based retrieval queries with controls for realism, diversity, and query formulation

Economic

🔎 **Query:** 설비투자와 건설 투자의 전기비 및 전년동기비 추이는 2023년 1분기까지 어떻게 달랐나요?

Relevant Pages



Energy

🔎 **Query:** 에너지사용량 신고제도와 에너지진단 제도 대상 절차 비교

Relevant Pages

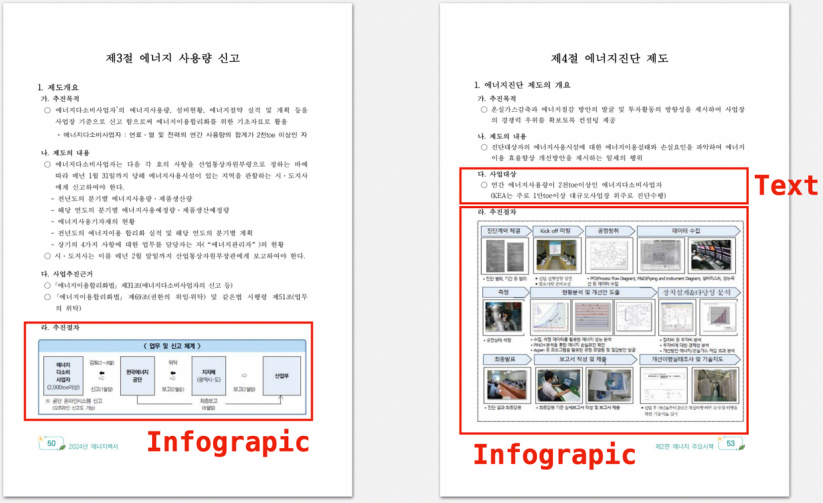
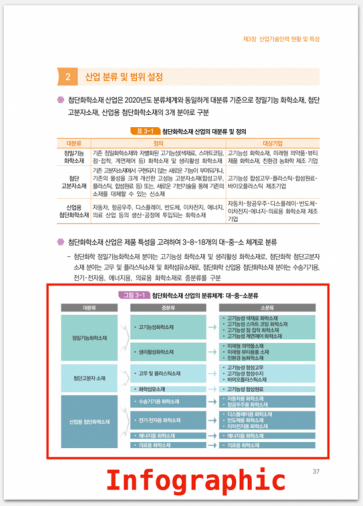


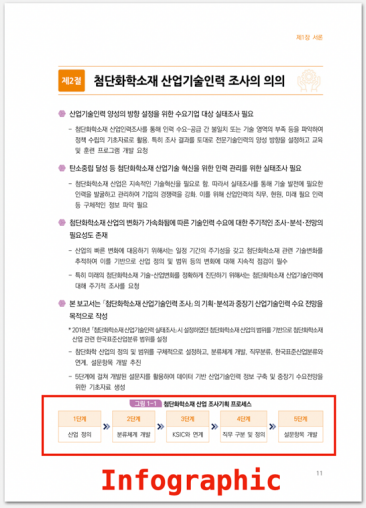
Figure 13: Example query-page pairs from the Economic and Energy subsets. Highlighted regions indicate the key evidence supporting each query.

HR

🔎 **Query:** 첨단화학소재의 5단계 조사 프로세스와 산업 분류 체계의 계층 구조를 분석하라.

Relevant Pages



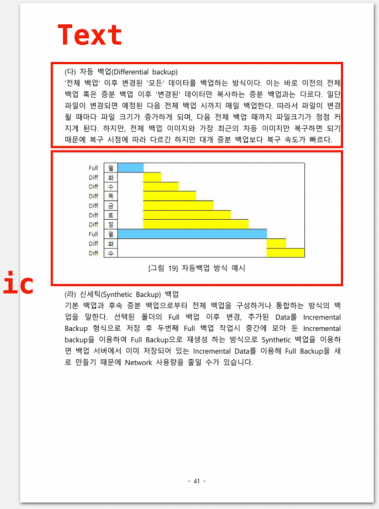


Cybersecurity

🔎 **Query:** 차등 백업과 중복제거 백업 방식의 저장 효율성과 복구 과정에서 어떤 차이가 있나요?

Relevant Pages

Text



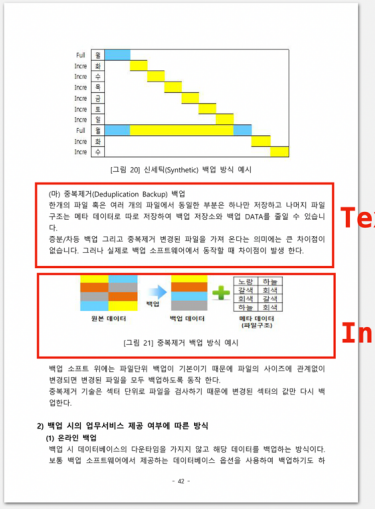
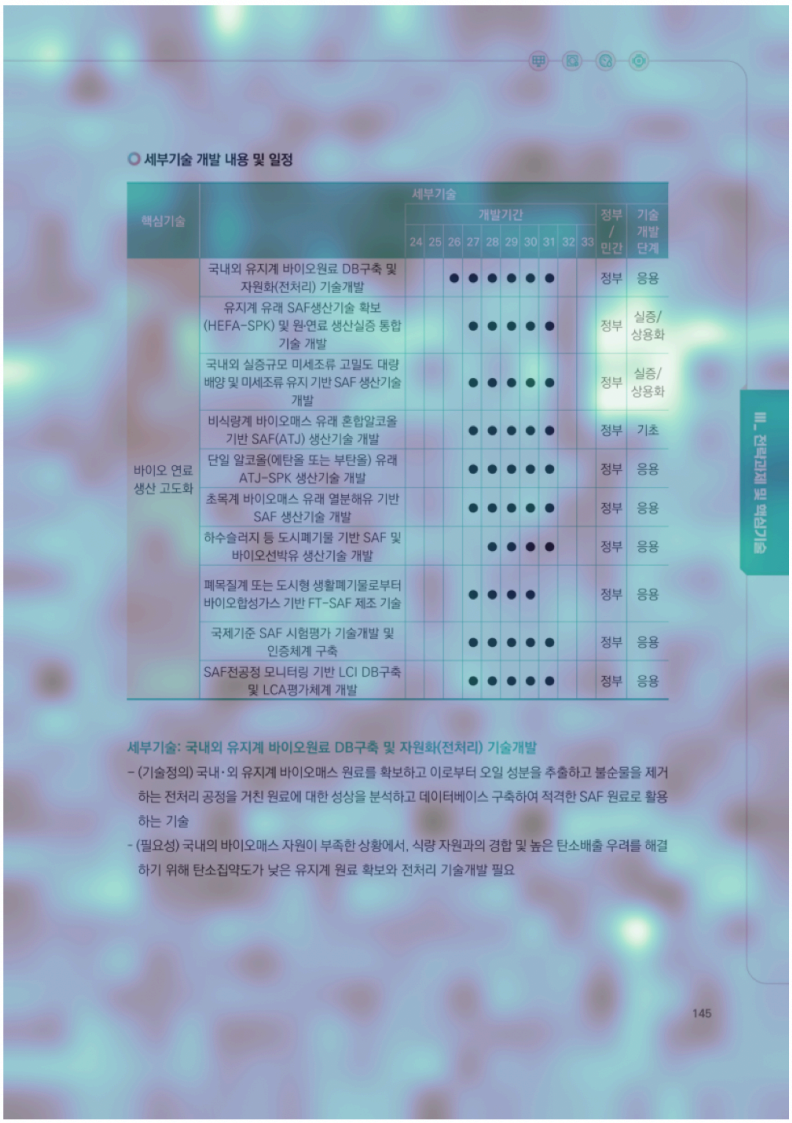


Figure 14: Example query-page pairs from the HR and Cybersecurity subsets. Highlighted regions indicate the key evidence supporting each query.



Query: **실증/상용화** 단계에 있는 SAF 생산기술의 총 개수는 몇 개인가요?

Figure 15: Additional query-to-document similarity heatmap for the fine-tuned `colqwen2-v1.0` model on a Korean document example. The model assigns high similarity to image patches corresponding to the term "실증/상용화," highlighting its focus on query-relevant textual evidence.

Title	Provider	Pages	License
Cybersecurity			
갠드크랩 랜섬웨어 악성코드 분석 기술 보고서	한국인터넷진흥원	92	No Restriction
사이버위협 동향보고서 (Windows 취약점 동향 및 업데이트 정책 등)	한국인터넷진흥원	104	No Restriction
사이버위협 동향보고서 (동형암호 기반 데이터 결합 및 분석 등)	한국인터넷진흥원	100	No Restriction
사이버 위협 동향보고서 (피싱 메일 공격 사례 등)	한국인터넷진흥원	96	No Restriction
사이버 위협 동향보고서 (기업 보안관리자의 크리덴셜 스터핑(Credential Stuffing) 공격 대응방안 등)	한국인터넷진흥원	104	No Restriction
사이버위협 동향보고서 (공인인증서 문제점과 DID 기술 전망 등)	한국인터넷진흥원	100	No Restriction
사이버위협 동향보고서 (ATT&CK Framework 개념과 이해 등)	한국인터넷진흥원	80	No Restriction
랜섬웨어 대응을 위한 안전한 정보시스템 백업 가이드(개정본)	한국인터넷진흥원	68	No Restriction
해킹진단도구 활용 사례 (취약한 관리자 계정 악용을 악용한 데이터 유출)	한국인터넷진흥원	23	No Restriction
해킹진단도구 활용 사례 (노출된 SMB 파일 서버를 통한 AD 환경 장악)	한국인터넷진흥원	23	No Restriction
해킹진단도구 활용 사례 (취약한 MS-SQL 서버를 통한 랜섬웨어 침해사고)	한국인터넷진흥원	14	No Restriction
해킹진단도구 활용 사례 (AD 환경에서의 RAT 악성코드 감염)	한국인터넷진흥원	16	No Restriction
AD서버 악용 내부망 랜섬웨어 유포 사례 분석	한국인터넷진흥원	31	No Restriction
Log4j 위협 대응 보고서	한국인터넷진흥원	35	No Restriction
NAS 보안 가이드	한국인터넷진흥원	161	No Restriction
TTPs2 스피어 피싱을 통한 공격망 구성 방식 분석	한국인터넷진흥원	79	No Restriction
TTPs3 공격자의 악성코드 활용 전략 분석	한국인터넷진흥원	27	No Restriction
Economic			
최근 경제동향 (2021. 3월호)	기획재정부	80	No Restriction
최근 경제동향 (2021. 6월호)	기획재정부	80	No Restriction
최근 경제동향 (2021. 9월호)	기획재정부	80	No Restriction
최근 경제동향 (2021. 12월호)	기획재정부	80	No Restriction
최근 경제동향 (2022. 3월호)	기획재정부	80	No Restriction
최근 경제동향 (2022. 6월호)	기획재정부	80	No Restriction
최근 경제동향 (2022. 9월호)	기획재정부	80	No Restriction
최근 경제동향 (2022. 12월호)	기획재정부	80	No Restriction
최근 경제동향 (2023. 3월호)	기획재정부	82	No Restriction
최근 경제동향 (2023. 6월호)	기획재정부	80	No Restriction
최근 경제동향 (2023. 9월호)	기획재정부	80	No Restriction
최근 경제동향 (2023. 12월호)	기획재정부	80	No Restriction
최근 경제동향 (2024. 3월호)	기획재정부	77	No Restriction
최근 경제동향 (2024. 6월호)	기획재정부	77	No Restriction
최근 경제동향 (2024. 9월호)	기획재정부	77	No Restriction
최근 경제동향 (2024. 12월호)	기획재정부	77	No Restriction
최근 경제동향 (2025. 3월호)	기획재정부	77	No Restriction
최근 경제동향 (2025. 6월호)	기획재정부	77	No Restriction
최근 경제동향 (2025. 9월호)	기획재정부	77	No Restriction
최근 경제동향 (2025. 12월호)	기획재정부	77	No Restriction
Energy			
해외전력산업동향 (2017 China)	한국전력거래소	57	No Restriction
해외전력산업동향 (2017 Japan)	한국전력거래소	43	No Restriction
해외전력산업동향 (2017 USA)	한국전력거래소	43	No Restriction
제4차 에너지기술개발계획 기술로드맵: 에너지저장	한국에너지기술평가원	172	KOGL Type 2
제4차 에너지기술개발계획 기술로드맵: 총괄	한국에너지기술평가원	76	KOGL Type 2
대전광역시 신재생에너지 보급계획	대전광역시	536	No Restriction
해외전력산업동향 (2023)	한국전력거래소	478	No Restriction
인천광역시 에너지백서	인천광역시	259	No Restriction
제5차 에너지기술개발계획 기술로드맵: 수요관리	한국에너지기술평가원	85	KOGL Type 2
제5차 에너지기술개발계획 기술로드맵: 효율향상	한국에너지기술평가원	162	KOGL Type 2
에너지 기술정책 포커스 (2025 주요국 기후에너지정책)	한국에너지기술연구원	122	No Restriction
HR			
고용형태별 근로실태조사 보고서	고용노동부	277	KOGL Type 1
블라인드 채용 가이드북	고용노동부	88	No Restriction
일 · 가정 양립 실태조사 보고서	고용노동부	423	No Restriction
한국직업전망	한국고용정보원	668	KOGL Type 2
유망 신산업 산업기술인력 전망: 이차전지	한국산업기술진흥원	134	No Restriction
유망 신산업 산업기술인력 전망: 첨단화학소재	한국산업기술진흥원	134	No Restriction
유망 신산업 산업기술인력 전망: 첨단섬유소재	한국산업기술진흥원	148	No Restriction
유망 신산업 산업기술인력 전망: 신금속소재	한국산업기술진흥원	136	No Restriction
유망 신산업 산업기술인력 전망: 차세대세라믹소재	한국산업기술진흥원	138	No Restriction

Table 7: Document metadata for KoViDoRe across all subsets.

Title	Provider	Pages	License
에너지총조사	기후에너지환경부	774	No Restrictions
주요업무계획	경기도 하남시	640	No Restrictions
지방공무원 인사실무	행정안전부	521	No Restrictions
항만편람	해양수산부	515	No Restrictions
국가연구개발사업 상위평가보고서	과학기술정보통신부	479	No Restrictions
연구보고서 현황	한국방송통신전파진흥원	425	No Restrictions
작업환경실태조사 보고서	한국산업안전보건공단	363	No Restrictions
국가연구개발사업 특정평가보고서	과학기술정보통신부	323	No Restrictions
관리형매립지 조사결과보고서	수도권매립지관리공사	285	No Restrictions
ICT 융복합 시설의 안전한 전자파 환경 기반 조성 연구	국립전파연구원	253	KOGL Type 1
전자파 흡수전력밀도 등 전자파 인체노출량 평가기술 연구	국립전파연구원	249	KOGL Type 1
해외건설 세무업무 매뉴얼	국토교통부	216	No Restrictions
처분시설 부지주변 방사선환경조사 보고서	한국원자력환경공단	211	KOGL Type 1
스마트 안전유지관리 시설물 확대방안 마련 용역 보고서	국토안전관리원	210	No Restrictions
해양수산발전기본계획	해양수산부	204	No Restrictions
디지털미디어 허브 조성을 위한 빛마루 중장기 전략 연구보고서	한국방송통신전파진흥원	195	No Restrictions
유엔개황 책자	외교부	178	No Restrictions
기업체노동비용조사 보고서	고용노동부	153	No Restrictions
(PDF)인삼재배전서	경상북도	151	No Restrictions
디지털미디어 신산업 진흥 방안 및 인력수급 기초조사에 관한 연구보고서	한국방송통신전파진흥원	146	No Restrictions
무인도서 100선	해양수산부	125	KOGL Type 1
환경관리해역 기본계획	해양수산부	125	No Restrictions
해외건설 법률컨설팅 사례	국토교통부	121	No Restrictions
국내외 온라인 동영상 미디어 콘텐츠 시장 전망 및 정책 추진방향 연구보고서	한국방송통신전파진흥원	115	No Restrictions
합성데이터 생성 활용 안내서	개인정보보호위원회	110	KOGL Type 1
환경관리해역 기본계획	해양수산부	90	No Restrictions
국가공무원인재개발원 교육운영계획	인사혁신처	85	No Restrictions
중소기업 경제동향 정보	중소벤처기업연구원	75	No Restrictions
생체정보 보호 안내서	개인정보보호위원회	73	KOGL Type 1
인재개발 종합계획	인사혁신처	68	No Restrictions
공공외교 기본계획	외교부	59	No Restrictions
관광실태조사 정보	부산광역시	58	No Restrictions
카지노 비즈니스와 제도	그랜드코리아레저(주)	57	No Restrictions
모바일 전자정부서비스 앱 소스코드 검증 가이드라인	행정안전부	51	No Restrictions
개인정보 유출 등 사고 대응 매뉴얼	개인정보보호위원회	48	KOGL Type 1
교통 리포트	서울특별시	42	No Restrictions
블랙잭 게임의 이해	그랜드코리아레저(주)	32	No Restrictions
개인정보 유출 신고 동향 및 예방 방법	한국인터넷진흥원	32	No Restrictions
i SMR 및 SSNC 설명자료	한국수력원자력(주)	27	No Restrictions
농지개량행위 신고 업무지침	농림축산식품부	24	No Restrictions
미디어이슈_광고요금제 도입을 앞둔 넷플릭스에 대한 인식 및 이용 조사	한국언론진흥재단	23	KOGL Type 1
위성전파 감시 정보	중앙전파관리소	19	KOGL Type 1
미디어이슈_이대남 현상에 대한 인식	한국언론진흥재단	19	KOGL Type 1
미디어이슈_코로나19 관련 정보 이용 및 인식 현황	한국언론진흥재단	19	KOGL Type 1
해외시장 신용위험 보고서	한국무역보험공사	18	No Restrictions
중남미 관련 보고서 (제약바이오)	외교부	12	No Restrictions
중남미 관련 보고서 (방위산업)	외교부	10	No Restrictions
노인 일자리 및 사회활동 지원사업 시행 20년의 성과와 발전과제	한국노인인력개발원	9	No Restrictions
전국 주배관망 가스인입지점별 인입가능량	한국가스공사	3	No Restrictions

Table 8: Document metadata for Ko-VDR Train Public.