ViDoRe V3: A Comprehensive Evaluation of Retrieval Augmented Generation in Complex Real-World Scenarios

António Loison[*†] **Quentin Macé**[*1,3] **Antoine Edy**[*1]
Victor Xing[1] **Tom Balough**[2] **Gabriel Moreira**[2] **Bo Liu**[2]
Manuel Faysse[3†] **Céline Hudelot**[3] **Gautier Viaud**[1]
[1]Illuin Technology [2]NVIDIA [3]CentraleSupélec, Paris-Saclay
{quentin.mace, antoine.edy, gautier.viaud}@illuin.tech[‡]

Abstract

Retrieval-Augmented Generation (RAG) pipelines must address challenges beyond simple single-document retrieval, such as interpreting visual elements (tables, charts, images), synthesizing information across documents, and providing accurate source grounding. Existing benchmarks fail to capture this complexity, often focusing on textual data, single-document comprehension, or evaluating retrieval and generation in isolation. We introduce **ViDoRe V3**, a comprehensive multi-modal RAG benchmark featuring multi-type queries over visually rich document corpora. It covers 10 datasets across diverse professional domains, comprising 26,000 document pages paired with 3,099 human-verified queries, each available in 6 languages. Through 12,000 hours of human annotation effort, we provide high-quality annotations for retrieval relevance, bounding box localization, and verified reference answers. Our evaluation of state-of-the-art RAG pipelines reveals that visual retrievers outperform textual ones, late-interaction models and textual reranking substantially improve performance, and hybrid or purely visual contexts enhance answer generation quality. However, current models still struggle with non-textual elements, open-ended queries, and fine-grained visual grounding. To encourage progress in addressing these challenges, the benchmark is released under a commercially permissive license[1].

1 Introduction

Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) has become the dominant paradigm for knowledge-intensive NLP tasks (Gao et al., 2024; Fan et al., 2024). Yet practical deployments introduce complexities that academic benchmarks often overlook when focusing on single-document textual retrieval. First, documents encode critical information in visual elements such as tables, charts, and images designed for human interpretation, which text-only pipelines often ignore (Abootorabi et al., 2025; Cho et al., 2024a). Second, user queries often require open-ended synthesis, comparison, and reasoning over scattered information, not simple factoid lookup (Tang and Yang, 2024; Conti et al., 2025; Thakur et al., 2025). Third, trustworthy systems must ground responses to specific source locations (e.g., bounding boxes), to mitigate hallucinations (Gao et al., 2023; Ma et al., 2024b).

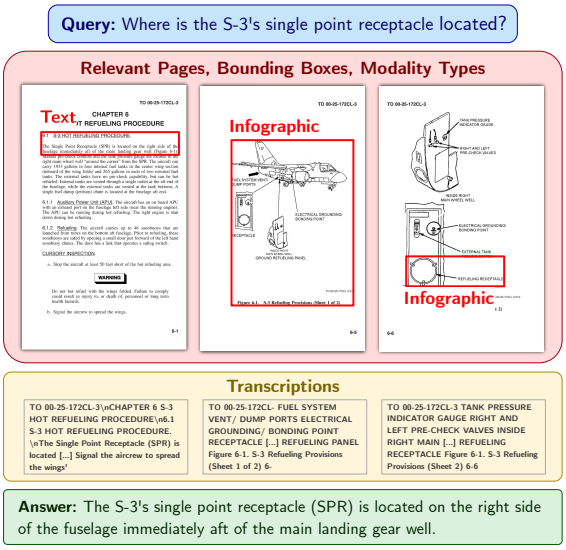


Figure 1: **ViDoRe V3 sample.** Each query is annotated with the relevant pages, a document-grounded answer, bounding boxes localizing supporting evidence and modality labels for each bounding box. Documents are provided in image, text and PDF formats.

Existing benchmarks leave these requirements only partially addressed. Early Visual Document Understanding (VDU) benchmarks focus on single-page comprehension, ignoring the complexity of large document corpora (Mathew et al., 2021b). Recent retrieval-centric benchmarks do not evalu-

[*]Equal contribution
[†]Work done while at Illuin Technology
[‡]Contact emails
[1]https://hf.co/vidore

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16570–16600

ate generation quality and grounding (Faysse et al., 2025; Günther et al., 2025). Some multimodal datasets attempt to bridge this gap but rely on extractive, short-answer tasks that fail to exercise complex reasoning (Cho et al., 2024b), or lack multilingual diversity and fine-grained visual grounding (Peng et al., 2025).

To address these limitations, we introduce **ViDoRe V3**, a benchmark designed for complex and realistic end-to-end RAG evaluation on visually rich document corpora. Our contributions are:

1. A Human Annotation Methodology for Realistic Queries We propose an annotation protocol for generating diverse queries and fine-grained query-page annotations. By restricting annotator access to document content during query formulation, we capture authentic search behaviors and mitigate bias toward simple extractive queries. Vision-Language Model (VLM) filtering combined with human expert verification enables efficient, high-quality annotation at scale.

2. The ViDoRe V3 Benchmark Applying this methodology to 10 industry-relevant document corpora, we build ViDoRe V3, a multilingual RAG benchmark comprising 26,000 pages and 3,099 queries, each available in 6 languages. Two datasets are held out as a private test set to mitigate overfitting. The benchmark is fully integrated into the MTEB ecosystem and leaderboard[2] (Muennighoff et al., 2023), and the public datasets are released under a commercially permissive license.

3. Comprehensive Evaluation and Insights Leveraging our granular annotations, we benchmark state-of-the-art models on **(i)** retrieval accuracy by modality and language, **(ii)** answer quality across diverse retrieval pipeline configurations, and **(iii)** visual grounding fidelity. Our analysis surfaces actionable findings for RAG practitioners.

2 Related Work

Component-Level Benchmarks (VDU and Retrieval) VDU has traditionally relied on single-page datasets like DocVQA (Mathew et al., 2021b), alongside domain-specialized variants (Mathew et al., 2021a; Zhu et al., 2022; Wang et al., 2024). These ignore the multi-page context inherent to RAG. Recent work evaluating bounding-box source grounding (Yu et al., 2025b) proposes single-page and multi-page tasks but does not address the retrieval component. Conversely, the emergence of late-interaction visual retrievers (Ma et al., 2024a; Faysse et al., 2025; Yu et al., 2025a; Xu et al., 2025) spurred the creation of retrieval-centric visual benchmarks like Jina-VDR (Günther et al., 2025) and ViDoRe V1&V2 (Faysse et al., 2025; Macé et al., 2025), but none of these benchmarks jointly evaluate retrieval and answer generation.

End-to-End Multimodal RAG While recent textual RAG benchmarks now capture complex user needs like reasoning or summarizing (Thakur et al., 2025; Tang and Yang, 2024; Su et al., 2024), multimodal evaluation often remains limited to single page queries (Faysse et al., 2025). Multi-page datasets like DUDE (Van Landeghem et al., 2023), M3DocRAG (Cho et al., 2024a), ViDoSeek (Wang et al., 2025) or Real-MM-RAG (Wasserman et al., 2025) prioritize extractive retrieval, lacking the diversity of queries encountered in realistic settings. UniDocBench (Peng et al., 2025) represents a concurrent effort that similarly addresses diverse query types and provides comparative evaluation across multiple RAG paradigms. While this benchmark offers valuable contributions, it relies on synthetically generated queries via knowledge-graph traversal, is restricted to English documents, and constrains grounding annotations to parsed document elements. In contrast, our benchmark offers several complementary strengths: fully human-verified annotations, a cross-lingual setup, free-form bounding box annotations, and a more systematic evaluation of individual visual RAG pipeline components.

3 Benchmark Creation

We design the benchmark to mirror the diversity of information retrieval situations in large-scale realistic environments. To enable pipeline-agnostic evaluation of the 3 core RAG components (retrieval, generation and grounding), while avoiding limitations of synthetic benchmarks, we employ a rigorous three-stage human-in-the-loop annotation process involving document collection, query generation and grounded query answering (Figure 2).

3.1 Document Collection

We curate 10 diverse corpora by manually selecting openly-licensed documents from governmental, educational, and industry sources, focusing on English and French documents (7 and 3 corpora respectively). The corpora span Finance, Computer Science, Energy, Pharmaceuticals, Human

[2] https://mteb-leaderboard.hf.space

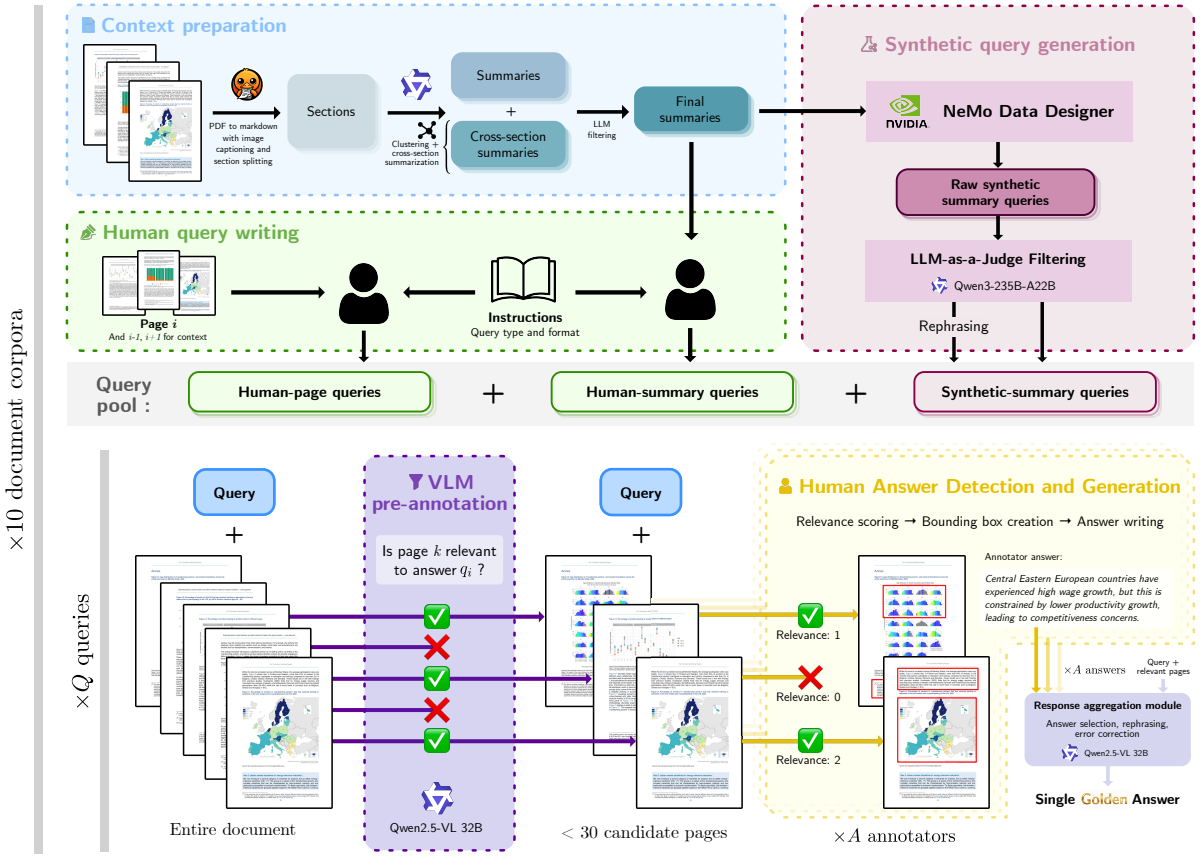


Figure 2: **Overview of the benchmark creation process.** Queries are sourced from 3 streams: *human extractive* (using raw pages), *human blind contextual* (using summaries to mitigate extractive bias), and *synthetic blind contextual*. For each query, a VLM pre-filtered subset of candidate pages is labeled by 1–3 human annotators that perform relevance scoring, bounding box localization and answer generation. A final response aggregation combines annotator answers into a single answer.

Resources, Industrial Maintenance, Telecom, and Physics. Each features domain-specific terminology and document structures representative of realistic retrieval tasks (details in Table 6).

3.2 Query Generation

Query Taxonomy To evaluate document visual retrieval systems across diverse realistic scenarios, we develop a query taxonomy with two orthogonal dimensions: *Query Type*, defining the user's information need, and *Query Format*, describing the query's syntactic structure. This dual-axis classification enables more nuanced performance analysis than benchmarks focusing solely on interrogative extractive queries. We define 7 Query Types: *open-ended*, *extractive*, *numerical*, *multi-hop*, *compare-contrast*, *boolean*, and *enumerative*, and 3 Query Formats: *question*, *keyword*, and *instruction*.

Context Preparation We further ensure query diversity by pulling summaries from a heterogeneous set of contexts during the generation process. Two types of input contexts are used: specific document sections that target local information retrieval and cross-section summaries that target multi-document context retrieval. These summaries are produced through a refined process inspired by ViDoRe V2 (Macé et al., 2025). First, the text is extracted from PDFs using Docling (Auer et al., 2024) along with image descriptions. Then, summaries are generated with Qwen3-235B-Instruct (Qwen Team, 2025) from each document section. They are clustered to group similar summaries together using Qwen3-Embedding-0.6B (Zhang et al., 2025) as embedder, UMAP (McInnes et al., 2020) for dimension reduction and HDBSCAN (Campello et al., 2013) for clustering. Additionally, cross-section summaries are produced by synthesizing the summaries of 2 to 3 randomly selected sections per cluster. From this pool of summaries, a final subset is curated to maintain a strict balance between single-section and cross-section summaries. The selection also

ensures an even distribution across section modalities (text, images, and tables) as defined by the Docling element classification.

Synthetic Query Generation Queries are generated from the summaries using a first synthetic generation pipeline based on Qwen3-235B. For each summary, a prompt is constructed by sampling a query type and format at random, together with variable attributes such as length and difficulty, in order to promote diversity. The generated queries are subsequently evaluated by the same LLM acting as an automatic judge, which filters outputs according to 4 criteria: information richness, domain relevance, clarity and adherence to query type/format. Finally, 50% of the retained queries are rephrased to further enhance linguistic variance. This pipeline is implemented using NeMo Data Designer (NeMo Data Designer Team, 2025) to facilitate generation scaling.

Human Query Writing Human annotators are provided 2 kinds of contexts: synthetic summaries or specific PDF pages. They are tasked with generating one query following a specific query type and format and one query of their choice that is most adapted to the context provided.

3.3 Answer Detection and Generation

Queries are filtered and linked to relevant pages using a hybrid pipeline of VLM pre-filtering and human annotation. It is followed by human answer annotation and visual grounding.

Query-Page Linking Given the scale of our corpora, manual verification of each page relevance for each query is intractable. We therefore adopt a two-stage annotation pipeline. First, Qwen2.5-VL-32B-Instruct (Bai et al., 2025) pre-filters candidate pages by assessing whether each page image is relevant to the query. Queries whose answers span more than 30 flagged pages are discarded. Human annotators then review the remaining query-page pairs, evaluating query quality and rating page relevance on a three-point scale (Not Relevant, Critically Relevant, Fully Relevant). We selected Qwen2.5-VL-32B-Instruct for its high recall, prioritizing coverage over precision and leaving final relevance judgments entirely to human annotators (see Appendix G for details and distributional validation).

Relevant Page Selection To ensure annotation quality, each task is completed by multiple annotators and reviewed by annotation supervisors. Since VLM pre-filtering biases the distribution toward relevant pages, we report Gwet's AC2, as it remains stable under prevalence skew, at 0.760 (see Section D for dataset-level breakdowns). Given this strong but imperfect agreement, we implement a tiered review process: extractive queries require at least one annotator and one reviewer, while more complex non-extractive queries require at least two annotators and one reviewer. A page is retained as relevant if marked by either (i) one annotator and one reviewer, or (ii) at least two annotators.

Answer Generation For each selected query, annotators were tasked with writing an answer based on the pages they marked as relevant. Given that different annotators might have different answer interpretations and tend not to be exhaustive in their answers, we use Qwen2.5-VL-32B-Instruct to generate a final answer based on the relevant page images marked by the annotators and their answers. To validate that this aggregation faithfully preserves annotator intent, we evaluated the VLM-aggregated answer against individual annotator responses using a GPT-5.2 judge to assess factual consistency. The aggregated answer matched the exact informational content (with minor paraphrasing) of at least one annotator's response in 86.3% of cases, confirming the VLM predominantly acts as a selector. For the remaining 13.7% of divergent cases, manual review of a random subset showed the aggregated answer was judged superior in most cases, either by merging complementary information from two incomplete responses or by correcting verifiable factual errors in individual annotations.

Bounding Boxes and Modality Types For each relevant page, annotators delineate bounding boxes around content supporting the query and attribute a modality type to each bounding box: Text, Table, Chart, Infographic, Image, Mixed or Other. Because multiple valid interpretations of bounding boxes can exist, we perform a consistency study to evaluate inter-annotator agreement and establish a human performance upper bound for the task.

We compute inter-annotator agreement on the subset of query-page pairs labeled by two or three annotators. For each annotator, we merge all their bounding boxes into a single zone. We then compare zones across annotators by measuring pixel-level overlap, reporting Intersection over Union (IoU) and F1 score (Dice coefficient). When three annotators label the same sample, we average over

all pairwise comparisons.

Across all 10 datasets, we observe an average IoU of 0.50 and F1 of 0.60. These moderate agreement scores reflect the inherent subjectivity of the task: annotators typically agreed on the relevant content but differed in granularity (Appendix J), with some marking tight bounds around specific content while others included surrounding context. Section 4.3 describes how our evaluation methodology accounts for this ambiguity and how model scores should be interpreted relative to this human ceiling.

Quality Control The annotation was conducted by a curated pool of 76 domain-qualified experts with native-level language proficiency. Quality control was performed by 13 senior annotators with enhanced domain knowledge and extensive annotation experience. Detailed protocols regarding the annotator pool and training are provided in Appendix C.

3.4 Final Query Distribution

We conducted a final human review to remove low-quality queries and resolve labeling ambiguities. Figure 3 shows the resulting distribution. Extractive queries predominate due to human annotator preference, followed by open-ended queries from targeted sampling. Multi-hop queries were the hardest to scale, suggesting a need for dedicated pipelines. Figure 4 details page modalities; while text is most prevalent, visual elements like tables, charts, and infographics are well-represented.

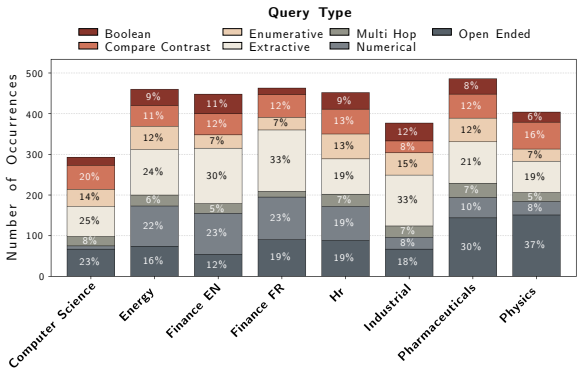


Figure 3: **Query Type Distribution per Domain**

3.5 Dataset Release and Distribution

We extend the benchmark to rigorously assess cross-lingual retrieval. While source documents are maintained in English and French, we use Qwen3-235B-Instruct to provide translations in 6 languages: English, French, Spanish, German, Italian, and Portuguese. This configuration challenges models to bridge the semantic gap between the query language and the document language, a critical requirement for modern RAG systems.

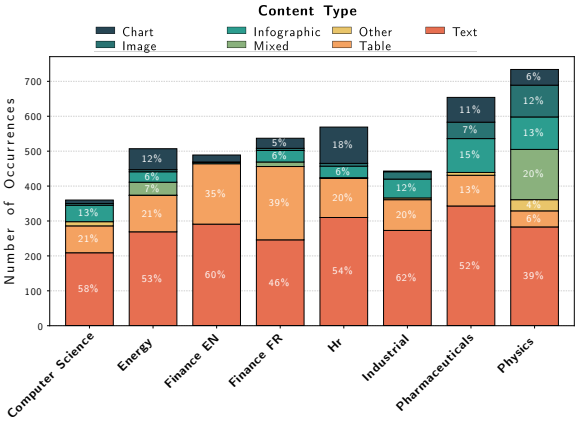


Figure 4: **Content Type Distribution per Domain**

Finally, to ensure the integrity of evaluation and mitigate data contamination (which was shown to be a major preoccupation for Information Retrieval (Liu et al., 2025)), we adopt a split-release strategy. 8 datasets are made public to facilitate research, while 2 are retained as private hold-out sets. This enables blind evaluation, ensuring that performance metrics reflect true generalization rather than overfitting to public samples.

4 Experiments and Results

Using our benchmark, we conduct extensive evaluations across all 3 components of RAG pipelines. We assess textual and visual retrievers and rerankers on retrieval performance, evaluate leading VLMs and LLMs on their ability to generate accurate answers from various retrieved contexts, and test VLMs on bounding box generation for visual grounding. From these results, we compile practical insights for RAG practitioners.

4.1 Retrieval

We evaluate a large panel of visual and textual retrievers on page-level retrieval ability. Visual retrievers are given page images, while textual retrievers process the Markdown text of each page processed by the NeMo Retriever extraction service[3] (NVIDIA Ingest Development Team,

[3]Chunking within pages or providing image descriptions did not improve our results. Thus, we report the results of the simplest pipeline.

Model	Size (B)	English Datasets							French Datasets			
		C.S.	Nucl.	Fin.	Phar.	H.R.	Ind.	Tel.	Phys.	Ener.	Fin.	**Avg.**
Textual Retrievers												
Qwen3-8B★	8	**71.7**	39.0	**49.4**	**59.2**	47.6	**40.4**	**62.8**	**45.6**	58.9	35.8	**51.0**
Jina-v4	3	64.3	**44.3**	48.4	54.9	**52.8**	38.4	56.3	43.6	**60.1**	**41.3**	50.4
LFM2-350M	0.35	63.5	37.8	39.0	56.4	43.5	34.4	56.9	41.8	47.0	28.2	44.9
Qwen3-0.6B★	0.6	66.4	32.8	42.7	50.6	37.7	31.6	55.7	43.3	51.3	25.8	43.8
BGE-M3★	0.57	58.0	30.2	39.8	52.0	42.4	28.5	51.6	35.9	49.8	25.2	41.3
BM25S	-	28.7	17.4	17.6	27.3	12.8	15.6	33.3	14.8	21.9	14.0	20.3
Visual Retrievers												
ColEmbed-3B-v2	3	**77.1**	**50.7**	**64.2**	**66.0**	**62.3**	**51.7**	**69.7**	47.0	64.9	44.4	**59.8**
Jina-v4	3	71.8	50.0	59.3	63.1	59.5	50.4	64.8	46.6	64.0	**46.1**	57.6
ColNomic-7B	7	76.2	45.0	56.6	62.3	58.7	50.1	67.2	**48.3**	64.0	45.5	57.4
ColEmbed-3B	3	75.2	49.1	60.9	63.7	58.7	47.1	67.0	45.1	62.1	43.8	57.3
ColNomic-3B	3	72.7	42.1	56.3	61.1	57.3	47.4	64.5	47.5	**65.0**	44.3	55.8
ColEmbed-1B	1	71.3	47.3	58.9	62.6	57.0	46.6	64.7	44.1	60.9	42.4	55.6
ColQwen2.5	3	72.3	38.1	52.3	57.9	51.2	41.3	61.3	45.9	59.7	39.1	51.9
Nomic-7B★	7	66.6	36.7	48.8	58.9	46.2	37.9	57.8	44.2	57.5	36.0	49.0
ColQwen2	2	68.6	35.7	39.0	52.2	45.1	38.3	57.4	41.6	48.8	20.0	44.7
Nomic-3B★	3	58.5	32.2	44.2	55.3	43.3	33.2	53.7	42.0	51.4	28.9	44.3
ColPali	7	65.3	32.9	34.4	53.1	44.8	35.6	54.0	41.7	47.1	21.8	43.1

Table 1: **Retrieval performance (NDCG@10) across the benchmark.** Best results per category in bold. ★: single-vector models. Following MTEB conventions, the average score is a macro-average over all datasets. Full model names and references are found in Table 8.

2024). The results reported in Table 1 corroborate findings from existing document retrieval benchmarks (Faysse et al., 2025; Günther et al., 2025): for a given parameter count, visual retrievers outperform textual retrievers, and late interaction methods score higher than dense methods.

We analyze ColEmbed-3B-v2, the best-performing retriever we evaluated across query type, content modality, and query language. A breakdown by query generation source is provided in Appendix E (Table 14).

Performance is aligned with query complexity Figure 5 shows that performance is inversely correlated with query complexity: simple query types such as Boolean and Numerical score significantly higher than Open-ended and Multi-hop queries. Question formulations consistently outperform Instruction and Keyword formats across nearly all categories, underscoring the need for improved handling of these query structures.

Visual Content and multi-page queries are hardest for retrievers Figure 6 highlights that queries involving visual content like tables or images tend to be more difficult. The Mixed content type scores the lowest, which suggests that integrating information across different modalities within a single page remains a challenge. Additionally, we observe a consistent decline in performance as the number of annotated pages increases (Figure 7), suggesting that retriever effectiveness decreases when aggregating information from multiple sources is required.

Figure 5: **ColEmbed-3B-v2 NDCG@10 by query type and format.**

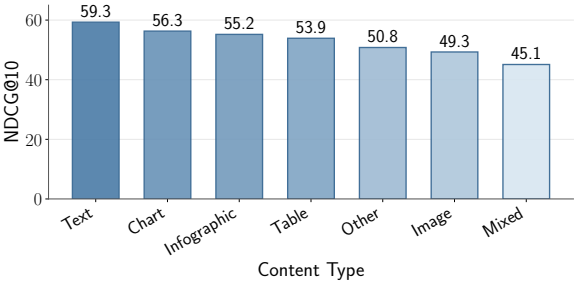


Figure 6: **ColEmbed-3B-v2 NDCG@10 by modality.**

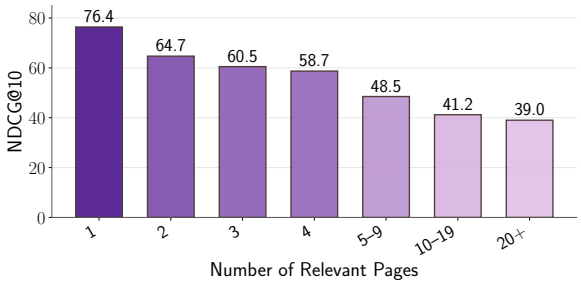


Figure 7: **ColEmbed-3B-v2 NDCG@10 by number of annotated pages.**

Cross-language queries degrade performance Retrieval performance is 2–3 points higher in monolingual settings (Table 9 and Table 10) than cross-lingual settings (Table 1), showing that models need to better adapt to these settings.

Textual rerankers outperform visual ones We evaluate the impact of adding a reranker to the textual and visual pipelines of the Jina-v4 retriever. We select zerank-2 (Zero Entropy, 2025) and jina-reranker-m0 (Jina AI, 2025) as two of the leading textual and visual rerankers to date. Results in Table 2 reveal a significant disparity in reranking efficacy between modalities. While the visual retriever initially outperforms the textual base, the textual reranker yields substantial gains (+13.2 NDCG@10), enabling the textual pipeline to achieve the highest overall retrieval performance. In contrast, the visual reranker provides only a marginal average improvement of +0.2 and degrades performance in 4 datasets, underscoring the need for better multilingual visual rerankers.

4.2 Final Answer Generation

We evaluate end-to-end answer quality by providing LLMs and VLMs with queries and their corresponding retrieved pages, examining the effects of retrieval pipeline selection, context modality, and generation model choice (Table 3). For this evaluation, we use the best-performing textual and visual retrieval pipelines. We additionally establish an upper bound using an *oracle* pipeline that supplies the model with ground-truth annotated pages.

In the *hybrid* configuration, we concatenate the top-5 results from the visual retriever (images) with the top-5 results from the textual retriever (text), without removing duplicates; the retrieval performance is detailed in Table 11. We also consider a hybrid *oracle* setup, which provides the model with all the ground-truth pages in both modalities.

The correctness of generated answers is assessed against the ground truth final answer by an LLM judge (details in Appendix I). Private datasets are omitted to maintain their integrity.

Some benchmark queries involve general knowledge manageable by LLMs without retrieval. To prevent memorization from confounding our assessment of the RAG pipeline, we stratify queries by difficulty based on parametric knowledge. A query is categorized as *easy* if any model in a 6-LLM panel answers it correctly without context; otherwise, it is labeled *hard*. Overall, 48.6 % of queries are easy (see Table 22 for details).

Visual context helps generation With a fixed Gemini 3 Pro generator, image-based context outperforms text-based context on the hard subset by 2.4 and 2.8 percentage points for the oracle and ColEmbed-3B-v2 pipelines, respectively (Table 3). This confirms that preserving the visual content of document pages provides better grounding for complex answer generation.

Hybrid retrieval yields the best performance on challenging queries The hybrid pipeline achieves 54.7 % accuracy on hard queries, surpassing both the strongest textual (52.1 %) and visual (54.5 %) baselines. This complementary effect suggests that text and image representations capture different aspects of document content, and their combination can provide more robust evidence for downstream generation.

Hard queries expose the limits of parametric knowledge in current models Even with oracle context, performance on hard queries lags behind easy queries by more than 10 percentage points. This gap suggests that the multi-step reasoning and long-context synthesis required for difficult queries remain challenging for current models. While the models we evaluate achieve comparable overall scores, their relative ranking may shift when parametric knowledge is less of an advantage, as shown by GPT 5.2 outperforming Gemini 3 Pro on easy queries but trailing on hard ones.

ViDoRe V3 leaves significant room for future retriever improvements The 10-point gap between the best non-oracle result (54.7 %) and the image oracle (64.7 %) on hard queries underscores substantial opportunities for improving the retrieval pipeline. Moreover, even with oracle contexts, Gemini 3 Pro performance remains modest, indi-

Model	English Datasets							French Datasets			Avg.
	C.S.	Nucl.	Fin.	Phar.	H.R.	Ind.	Tel.	Phys.	Ener.	Fin.	
Textual pipeline											
Jina-v4$^{\text{textual}}$	64.3	44.3	48.4	54.9	52.8	38.4	56.3	43.6	60.1	41.3	50.4
+ zerank-2	**82.1** ↑17.8	**53.5** ↑9.2	**69.2** ↑20.8	**66.2** ↑11.3	**66.5** ↑13.7	**53.2** ↑14.8	**71.5** ↑15.2	**48.2** ↑4.6	**71.5** ↑11.4	**53.7** ↑12.4	**63.6** ↑13.2
Visual pipeline											
Jina-v4$^{\text{visual}}$	71.8	50.0	59.3	63.1	59.5	50.4	64.8	46.6	64.0	46.1	57.6
+ jina-reranker-m0	76.7 ↑4.9	50.8 ↑0.8	59.2 ↓0.1	65.4 ↑2.3	56.0 ↓3.5	50.9 ↑0.5	70.8 ↑6.0	46.9 ↑0.3	61.7 ↓2.3	39.8 ↓6.3	57.8 ↑0.2

Table 2: **Retrieval performance (NDCG@10) of retriever + reranker pipelines.**

Retrieval pipeline	Context modality	Generation model	English Datasets					French Datasets			Avg. Hard	Avg. Easy	Avg. Global
			C.S.	Fin.	Phar.	H.R.	Ind.	Phys.	Ener.	Fin.			
Oracle	Text	Gemini 3 Pro	80.9	70.2	71.4	**72.3**	66.4	71.2	69.2	62.8	62.3	79.3	70.6
	Image		**86.5**	**70.6**	**76.1**	71.1	**68.2**	74.5	**69.8**	**64.1**	**64.7**	**79.7**	**72.6**
	Hybrid		86.0	68.9	73.4	70.4	65.4	69.2	69.5	62.8	63.4	77.5	70.7
Jina-v4$^{\text{text.}}$ + zerank-2	Text	Gemini 3 Pro	80.9	66.0	59.9	63.2	60.4	69.2	64.9	54.7	52.1	75.5	64.9
Jina-v4$^{\text{text.}}$ + zerank-2 & ColEmbed-3B-v2	Hybrid	Gemini 3 Pro	85.1	65.0	65.9	64.8	59.4	69.9	62.7	52.8	54.7	76.6	65.7
ColEmbed-3B-v2	Text	Gemini 3 Pro	82.3	62.5	61.0	62.9	56.2	64.9	62.3	49.4	51.7	73.0	62.7
		Kimi K2	81.4	56.6	59.1	55.7	55.8	73.8	60.4	43.1	44.6	74.3	60.7
	Image	Gemini 3 Pro	83.3	67.3	62.9	65.4	57.2	67.9	64.3	47.8	54.5	74.1	64.5
		Gemini 3 Flash	80.9	64.1	63.5	63.8	55.1	68.2	63.3	47.8	50.3	74.4	63.3
		GPT-5.2	**86.5**	59.5	68.1	66.0	61.5	**76.5**	66.2	49.1	54.1	78.1	66.7
		Qwen3-VL-235B	86.0	59.9	64.0	60.7	57.2	71.9	59.7	44.4	51.0	74.1	63.0

Table 3: **End-to-end evaluation of final answer generation.** We report the percentage of correct final answers as determined by an LLM judge across the 8 public datasets. "Oracle" rows represent the upper-bound performance using gold-standard contexts. Average Easy and Average Hard denote performance stratified by query difficulty. For each column, the best result is **bolded** and the best non-oracle result is underlined.

cating that generation models still struggle to fully exploit the provided information.

4.3 Visual Grounding

Beyond generating correct answers, it is highly desirable for RAG pipelines to identify where in the source documents the answer originates, enabling users to verify the grounding of the query answer. We therefore evaluate the ability of LLMs to generate accurate bounding boxes within their final answer. Among the few LLM families with visual grounding capabilities, we select Qwen3-VL-30B-A3B-Instruct and Gemini 3 Pro for evaluation. For each query, we provide the model with the candidate pages shown to the human annotators and prompt it to answer the query while inserting inline bounding boxes in XML format `<bboxes image="N"> ... </bboxes>` to delimit relevant content (full instructions in Appendix H).

We use the bounding boxes produced by the human annotators as our ground truth. Since each query may have 1–3 human annotators, we evaluate VLM predictions independently against each annotator using the same zone-based methodology as the inter-annotator consistency analysis (Section 3.3), and report the highest F1 score. This best-match strategy reflects the inherent subjectivity of evidence selection: annotators may legitimately highlight different regions to support the same answer, and a model should not be penalized for matching any valid interpretation.

Visual grounding lags human performance Inter-annotator agreement on evidence localization reaches an F1 of 0.602, whereas the best-performing models achieve markedly lower scores: 0.089 for Qwen3-VL-30B-A3B-Instruct and 0.065 for Gemini 3 Pro, underlining substantial room for improvement on this task. A page-level analysis (Table 4) reveals that on pages where humans provided bounding boxes, both models annotated the same page only 16–17 % of the time, while 26–27 % of human-annotated pages received no model annotation at all, highlighting recall as the primary bottleneck. Per-domain results and qualitative analysis appear in Appendix H and J.

Category	Outcome	Qwen3-VL-30B-A3B	Gemini 3 Pro
Agreement	Both annotated	17 %	16 %
	Neither annotated	46 %	49 %
Discrepancy	Model only	10 %	7 %
	Human only	26 %	27 %

Table 4: **Page-level bounding box agreement between models and human annotators.** Each page is classified by whether the model and human both annotated it, both left it unannotated, or only one provided annotations.

5 Conclusion

This work introduces ViDoRe V3, a human-annotated RAG benchmark that evaluates cross-lingual retrieval, final answer generation, and visual grounding on large industry-relevant document corpora. We design a human-in-the-loop annotation methodology, deployed in a 12,000-hour annotation campaign, that produces diverse realistic queries paired with relevant pages, bounding boxes, and reference answers. Evaluating state-of-the-art RAG pipelines, we find that visual retrievers outperform textual ones, late interaction and textual reranking yield substantial gains, and visual context improves answer generation quality. Looking ahead, ViDoRe V3 highlights several concrete research directions for practical multimodal RAG. Retriever models still struggle on cross-lingual and open-ended queries requiring visual interpretation, while VLMs need improvement in answer generation from multi-page contexts as well as accurate visual grounding. By providing a rigorous framework for evaluating these limitations, ViDoRe V3 serves as a catalyst for the development of more robust, intelligent document understanding models.

Limitations

Language coverage While our benchmark is multilingual, it is restricted to English and French source documents and queries in 6 high-resource Western European languages. Future iterations of the benchmark should include a more diverse set of language families and non-Latin scripts to mitigate this bias.

Document distribution bias Our benchmark focuses on publicly available long-form document corpora, representing one specific mode of existing document distribution. For example, enterprise RAG may need to handle a wider variety of document types, often in private repositories, that include noisy, short-form types such as emails, support tickets, or scanned handwritten notes that are not represented in our source documents.

Human annotation Annotations for open-ended reasoning and visual grounding inherently contain a degree of subjectivity. We acknowledge that for complex exploratory queries, multiple valid retrieval paths and answer formulations may exist outside of our annotated ground truths.

Ethical considerations

Annotator Welfare and Compensation. Human annotation was conducted by the creators of the benchmark and a single external annotation vendor. Multiple established vendors were evaluated with respect to the annotation protocol and relevant ethical considerations, and one vendor was selected based on demonstrated compliance with these criteria. Annotators were recruited from the vendor's existing workforce in accordance with the demographic requirements described in the Annotator Pool and Selection section (Section C) and were compensated at rates designed to provide fair pay based on geographic location and required skill sets. The data were curated such that annotators were not exposed to harmful or offensive content during the annotation process. The use of human annotators was limited to standard annotation and verification tasks for benchmark construction and did not constitute human-subjects research; accordingly, the data collection protocol was determined to be exempt from formal ethics review.

Data Licensing and Privacy. All documents included in the benchmark were manually selected from governmental, educational, and enterprise websites that met open license criteria. The annotations were collected in order not to contain any private or personally identifiable information and are GDPR-compliant. The benchmark is released under a commercially permissive license to facilitate broad research adoption while respecting the intellectual property rights of original document creators.

Linguistic and Geographic Bias. We acknowledge that our benchmark is restricted to English and French source documents and queries in 6 high-resource Western European languages. This limitation may inadvertently favor RAG systems optimized for these languages and does not reflect the full diversity of practical document retrieval scenarios globally. We encourage future work to extend

evaluation to underrepresented language families and non-Latin scripts.

Environmental Impact. The creation of this benchmark required substantial computational resources for VLM pre-filtering, synthetic query generation, and model evaluation. We report these costs to promote transparency: approximately 12,000 hours of human annotation effort and extensive GPU compute for model inference across our evaluation suite. Specifically, the compute totaled 3,000 hours on NVIDIA H100 GPUs on a low emission energy grid, with an estimated environmental impact of 200 kg CO_2e.

Acknowledgments

This work was conducted with contributions from NVIDIA. We thank all the people that allowed this work to happen, in particular Eric Tramel, Benedikt Schifferer, Mengyao Xu and Radek Osmulski, Erin Potter and Hannah Brandon. Crucially, we thank the dedicated team of annotators for their essential efforts.

It was carried out within the framework of the LIAGORA "LabCom", a joint laboratory supported by the French National Research Agency (ANR) and established between ILLUIN Technology and the MICS laboratory of CentraleSupelec. The benchmark was partially created using HPC resources from IDRIS with grant AD011016393.

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. *Preprint*, arXiv:2502.08826.

Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, and 1 others. 2024. Docling technical report. *arXiv preprint arXiv:2408.09869*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Antoine Chaffin. 2025. Gte-moderncolbert.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *arXiv preprint*. Version Number: 3.

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

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

Max Conti, Manuel Faysse, Gautier Viaud, Antoine Bosselut, Céline Hudelot, and Pierre Colombo. 2025. Context is gold to find the gold passage: Evaluating and training contextual document embeddings. *Preprint*, arXiv:2505.24782.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. *Preprint*, arXiv:2405.06211.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. *Preprint*, arXiv:2407.01449.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *Preprint*, arXiv:2305.14627.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Sedigheh Eslami, Scott Martens, Bo Wang, Nan Wang, and Han Xiao. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. *Preprint*, arXiv:2506.18902.

Jina AI. 2025. jina-reranker-m0: Multilingual multimodal document reranker. Accessed: 2025-12-22.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Liquid AI. 2025. Lfm2 technical report. *arXiv preprint arXiv:2511.23404*.

Frank Liu, Kenneth Enevoldsen, Roman Solomatin, Isaac Chung, Tom Aarsen, and Zoltán Fődi. 2025. Introducing rteb: A new standard for retrieval evaluation.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. *Preprint*, arXiv:2407.03618.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024a. Unifying multimodal retrieval via document screenshot embedding. *Preprint*, arXiv:2406.11251.

Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. 2024b. Visa: Retrieval augmented generation with visual source attribution. *Preprint*, arXiv:2412.14457.

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

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models. *Preprint*, arXiv:2504.05299.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. 2021a. Infographicvqa. *Preprint*, arXiv:2104.12756.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021b. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

NeMo Data Designer Team. 2025. Nemo data designer: A framework for generating synthetic data from scratch or based on your own seed data. `https://github.com/NVIDIA-NeMo/DataDesigner`. GitHub Repository.

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

NVIDIA Ingest Development Team. 2024. *NVIDIA Ingest: An accelerated pipeline for document ingestion*.

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

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, and 1 others. 2024. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*.

Rikiya Takehi, Benjamin Clavié, Sean Lee, and Aamir Shakir. 2025. Fantastic (small) retrievers and how to train them: mxbai-edge-colbert-v0 tech report. *Preprint*, arXiv:2510.14880.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *Preprint*, arXiv:2401.15391.

Paul Teiletche, Quentin Macé, Max Conti, Antonio Loison, Gautier Viaud, Pierre Colombo, and Manuel Faysse. 2025. Modernvbert: Towards smaller visual document retrievers. *arXiv preprint arXiv:2510.01149*.

Nandan Thakur, Jimmy Lin, Sam Havens, Michael Carbin, Omar Khattab, and Andrew Drozdov. 2025. Freshstack: Building realistic benchmarks for evaluating retrieval on technical documents. *Preprint*, arXiv:2504.13128.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, and 1 others. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540.

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. *arXiv preprint arXiv:2502.18017*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

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. *arXiv preprint arXiv:2502.12342*.

Mengyao Xu, Gabriel Moreira, Ronay Ak, Radek Osmulski, Yauhen Babakhin, Zhiding Yu, Benedikt Schifferer, and Even Oldridge. 2025. Llama nemoretriever colembed: Top-performing text-image retrieval model. *Preprint*, arXiv:2507.05513.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025a. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *Preprint*, arXiv:2410.10594.

Wenhan Yu, Wang Chen, Guanqiang Qi, Weikang Li, Yang Li, Lei Sha, Deguo Xia, and Jizhou Huang. 2025b. Bbox docvqa: A large scale bounding box grounded dataset for enhancing reasoning in document visual question answer. *Preprint*, arXiv:2511.15090.

Zero Entropy. 2025. Introducing zerank-2. Accessed: 2025-12-22.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866.

A Dataset examples

Domain: **Pharmaceuticals**
Query id: **17**
Query type: **Extractive**
Query format: **Question**

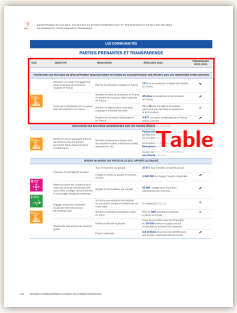
Query: According to FDA guidelines, what is the required format for displaying the strength of a small volume parenteral drug on its label?

Answer: According to FDA guidelines, the strength of a small-volume parenteral drug must be displayed as the quantity per total volume (e.g., 2 g/20 mL), immediately followed in parentheses by the quantity per milliliter (e.g., 100 mg/mL).

Domain: **Finance (FR)**
Query id: **32**
Query types: **Extractive, numerical**
Query format: **Keyword**

Query: taux de produits fabriqués en France Hermès 2023

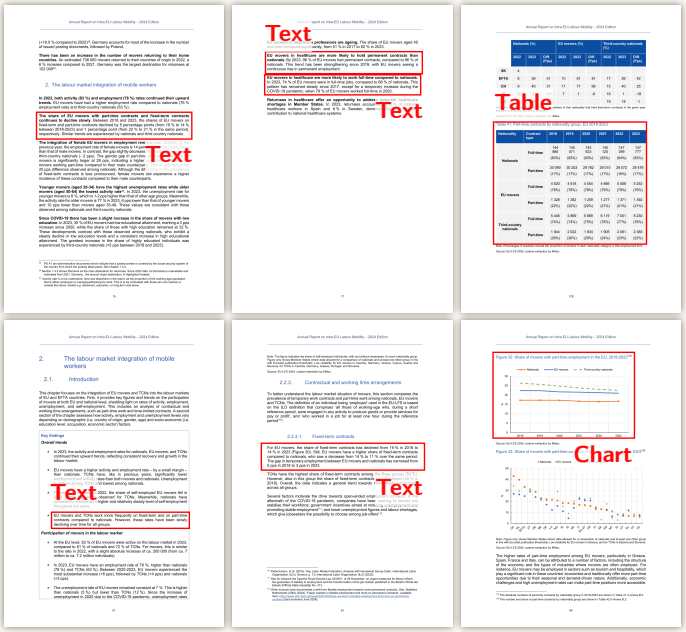
Answer: Le taux de produits d'Hermès fabriqués en France est de 74 % en 2023.



Domain: **Human Resources**
Query id: **0**
Query type: **Multi-hop**
Query format: **Question**

Query: What impact did the shifts in temporary contract usage from 2018 to 2023 have on the difference in employment stability between mobile EU workers and citizens of the host countries?

Answer: From 2018 to 2023, the job stability gap between mobile EU workers and nationals narrowed. The use of temporary contracts declined more for EU movers (from 19% to 14%) than for nationals (14% to 11%), reducing the difference between the two groups from 5 to 3 percentage points.



B Supplementary benchmark details

Domains Table 6 details the type of documents used in each corpus as well as several statistics.

Query type and format descriptions Table 5 describes the types and formats of the queries, while Figure 9 gives details about query type intersection frequency.

Query type by generation method Query type distributions by generation method (Figure 10) confirm that open-ended queries dominate synthetic queries as the synthetic pipeline attributed more weight to this type, while extractive queries dominate human-image queries since they are more naturally chosen by annotators.

C Annotator pool and training details

Annotator Pool and Selection. Annotation was conducted by a curated pool of 76 annotators who were selected based on having: (1) a bachelor's degree or higher in the relevant domain, (2) professional experience in the domain, (3) native-level language proficiency as required by task, and (4) prior experience with RAG, retrieval, or VQA annotation projects. Quality control was performed by 13 senior annotators with enhanced domain knowledge and extensive annotation experience, with project oversight provided by data leads with multiple years of experience in human data generation.

Training and Pilot Phase. The annotation process began with a comprehensive onboarding phase where annotators received task-specific training using gold-standard examples. For each domain, a pilot of several hundred tasks was conducted with 100% quality control coverage and multiple annotators per task. During this phase, data leads and the research team continuously evaluated annotations, provided clarifications, and refined guidelines. Inter-annotator agreement and time-per-task baselines were calculated to establish ongoing evaluation benchmarks. The pilot concluded upon validation of both data quality and guideline effectiveness.

D Supplementary agreement metrics

Pages were pre-filtered by a VLM before human annotation; as most pages shown to annotators were likely relevant, this created a skewed class distribution. This prevalence imbalance causes traditional chance-corrected metrics like Krippendorff's Alpha to appear paradoxically low even when annotators genuinely agree, as inflated expected chance agreement penalizes the score. To address this, we report 2 complementary metrics: Krippendorff's Alpha (ordinal) as the standard measure and Gwet's AC2 which remains stable under prevalence skew. Overall, annotators achieved $\alpha = 0.469$, AC2 $= 0.760$. The divergence between Alpha and AC2/Weighted Agreement is expected given the pre-filtered data and confirms substantial agreement despite the skewed distribution.

E Supplementary retrieval details

Retriever model reference Table 8 lists the retriever models evaluated in this work, along with their HuggingFace model names and citations.

Monolingual performance Tables 9 and 10 present the monolingual performance of our models, where retrieval is conducted using language-matched queries and documents for English and French, respectively.

Dataset	α (ord)	Gwet's AC2
Computer science	0.467	0.809
Energy	0.463	0.714
Finance (EN)	0.514	0.798
Finance (FR)	0.320	0.736
H.R.	0.413	0.793
Industrial Maintenance	0.496	0.740
Telecom	0.464	0.772
Nuclear	0.389	0.794
Pharma	0.478	0.755
Physics	0.213	0.334
Overall	**0.469**	**0.760**

Table 7: Inter-annotator agreement for relevance ratings by dataset.

Additional Retrieval Modality Performances To evaluate the hybrid retrieval setup, we use the multimodal Jina-v4 model to generate separate visual and textual rankings. We then construct a hybrid retrieval set by merging the top-5 results from each modality and removing duplicates. Because this set-union operation does not preserve a strict ranking order, we report the unranked F1 score. As shown in Table 11, the hybrid approach consistently outperforms single-modality baselines.

Category	Definition	Example
	Query Types	
Open-ended	Seeks explanatory or descriptive information that requires synthesis.	What drives the rise in women's workforce involvement in EU nations?
Extractive	Requires the retrieval of a specific piece of information.	Bank of America preferred stock MM dividend rate
Compare Contrast	Mandates a comparison between multiple entities or data points.	Explain the factors contributing to the reduction in R2R rates for ANDAs.
Boolean	Poses a yes/no question necessitating multi-step reasoning.	Did JPMorganChase execute more than half of its planned repurchase program?
Numerical	Asks for a specific quantitative value that must be derived or calculated.	percentage increase in Morgan Stanley revenue from 2023 to 2024
Multi-hop	Requires integrating information from multiple sections or sources.	Summarize the steps involved in error reporting in ISMP's MERP.
Enumerative	Requests a list of all instances sharing a common property.	Specify the ISCO codes used to define domestic workers in the EU.
	Query Formats	
Question	An interrogative sentence.	What was Citigroup's net interest margin in 2024?
Keyword	A non-verbal phrase or set of terms.	female employment rate European Union 2023
Instruction	A directive specifying a task.	Identify the use case of a drill point gauge.

Table 5: Taxonomy of Query Types and Formats.

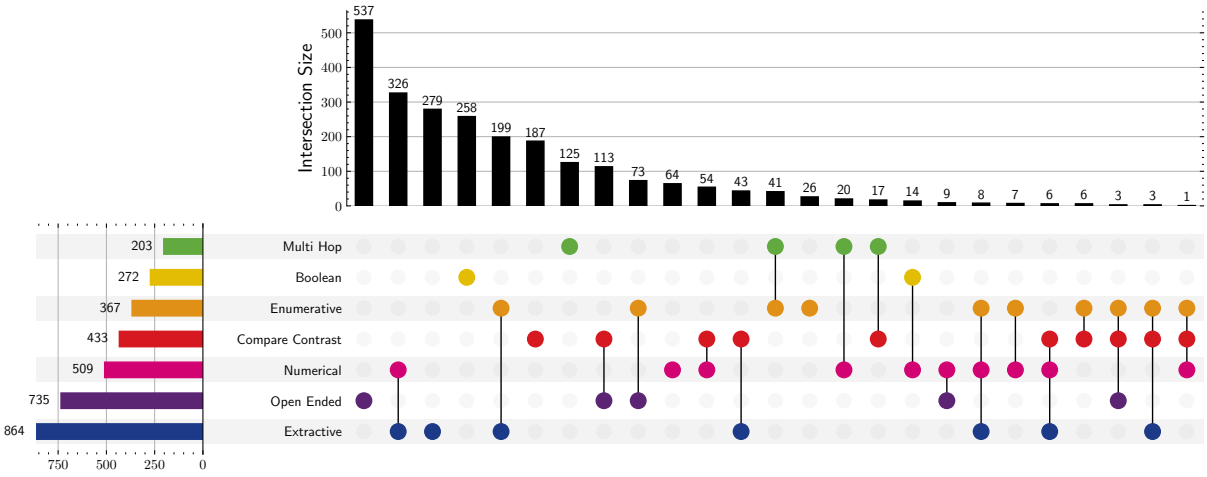


Figure 9: **UpSet plot illustrating the distribution and intersection of query types in ViDoRe V3.** The horizontal bars on the left display the total count of queries for each individual type. The vertical bars at the top represent the frequency of specific combinations (intersections), as indicated by the connected dots in the matrix below. While *Extractive* queries are the most prevalent overall, *Open Ended* queries form the dominant unique category. Complex dependencies are evident in the frequent intersection of *Enumerative* and *Extractive types*, indicating a substantial subset of queries requiring list-based fact retrieval.

F ColEmbed-3B-v2 performance breakdown

Table 12 details the retrieval scores of ColEmbed-3B-v2 by query language, highlighting small performance variations by language.

Performance by number of annotated pages As seen in Figure 7, performance drops with the number of annotated pages. However, a potential confounding factor is the correlation between query type and the number of annotated pages, since more complex query types also have higher number of annotated pages (Figure 11). We perform a stratified regression analysis to isolate these two effects.

We model NDCG@10 as a linear function of the number of annotated pages (P) stratified by query type. For each of the 7 query types, we fit an ordinary least squares regression:

$$NDCG@10 = a \cdot P + b + \epsilon.$$

Results in Figure 12 and Table 13 reveal that all query types suffer a significant performance

Corpus	Domain(s)	Description	Lang.	# Docs	# Pages	# Queries*	Main modalities
U.S. Public Company Annual Reports	Finance-EN	Consists of 6 10-K annual reports from major U.S. financial institutions for the fiscal year ended December 31, 2024.	en	6	2942	309	Text, Table
Computer Science Textbooks	Computer Science / Education	Consists of two open-source, peer-reviewed textbooks from OpenStax covering foundational topics in computer science, Python, and data science.	en	2	1360	215	Books
FDA Reports	Pharmaceuticals	Consists of FDA presentations and Springer books (2016–2023) covering regulatory policies, drug development, and public health initiatives.	en	52	2313	364	Slides, Books
HR Reports from EU	HR	Includes recent European Commission reports and papers on EU labour markets, social development, and employment policies.	en	14	1110	318	Reports
USAF Technical Orders	Industrial Maintenance	Comprises U.S. military technical orders and manuals for aircraft maintenance, safety procedures, and material handling, revised through 2025.	en	27	5244	283	Manuals
French Physics Lectures	Physics	A collection of educational materials offering an interdisciplinary exploration of modern physics and complexity science.	fr	42	1674	302	Slides
French Public Company Annual Reports	Finance-FR	Contains the 2023–2024 annual reports of major French luxury companies (Dior, Hermès, Kering, L'Oréal, LVMH).	fr	5	2384	320	Reports
French Governmental Energy Reports	Energy	Gathers official documents from French public agencies on energy, economic, and environmental issues in France.	fr	42	2229	308	Reports, Slides

Table 6: Description of ViDoRe V3 public corpora. *Number of queries is without translations

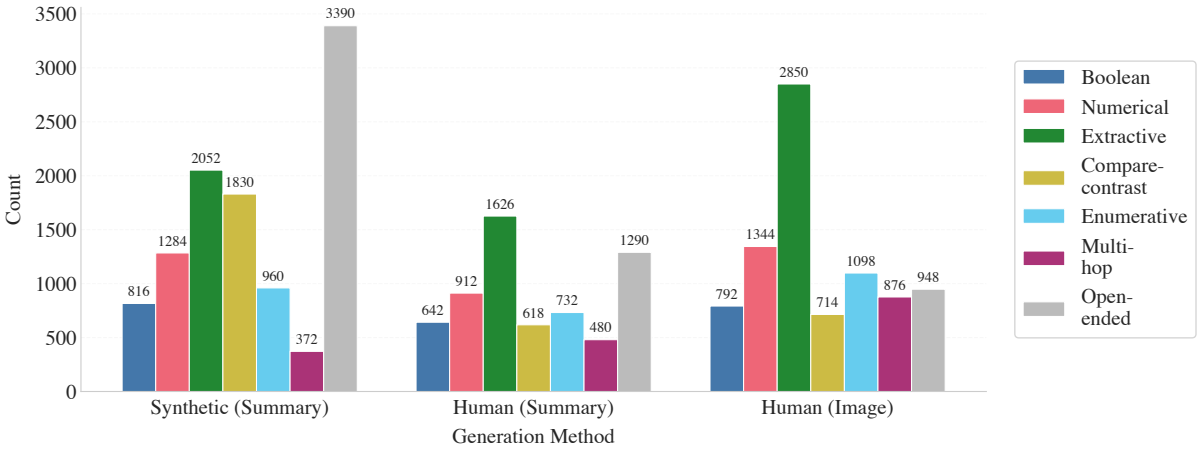


Figure 10: **Query type distribution by generation method.**

Model alias	Full model name	Reference
Qwen3-8B	Qwen3-Embedding-8B	Zhang et al. (2025)
Jina-v4	jina-embeddings-v4	Günther et al. (2025)
Qwen3-0.6B	Qwen3-Embedding-0.6B	Zhang et al. (2025)
LFM2-350M	LFM2-ColBERT-350M	Liquid AI (2025)
BGE-M3	BGE-M3	Chen et al. (2024)
BM25S	BM25S	Lù (2024)
ColEmbed-3B-v2	llama-nemoretriever-colembed-3b-v2	Xu et al. (2025)
ColNomic-7B	colnomic-embed-multimodal-7b	Nomic Team (2025)
ColEmbed-3B	llama-nemoretriever-colembed-3b-v1	Xu et al. (2025)
ColNomic-3B	colnomic-embed-multimodal-3b	Nomic Team (2025)
ColEmbed-1B	llama-nemoretriever-colembed-1b-v1	Xu et al. (2025)
ColQwen2.5	colqwen2.5-v0.2	Faysse et al. (2025)
Nomic-7B	nomic-embed-multimodal-7b	Nomic Team (2025)
ColQwen2	colqwen2-v1.0	Faysse et al. (2025)
Nomic-3B	nomic-embed-multimodal-7b	Nomic Team (2025)
ColPali	colpali-v1.3	Faysse et al. (2025)
Mxbai Edge 32M	mxbai-edge-colbert-v0-32m	Takehi et al. (2025)
GTE-ModernColBERT	GTE-ModernColBERT-v1	Chaffin (2025)
ColModernVBERT	colmodernvbert	Teiletche et al. (2025)
ColSmol-256M	colSmol-256M	Marafioti et al. (2025)

Table 8: **Retriever reference table.** Model aliases used in Tables 1, 9, and 10 are mapped to their HuggingFace model name and citation.

Figure 11: **Average number of annotated pages by query type.**

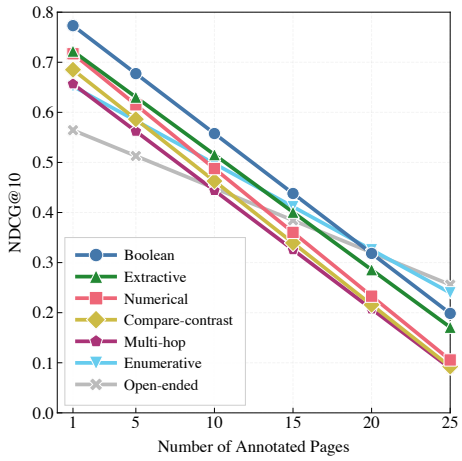


Figure 12: **ColEmbed-3B-v2 NDCG@10 by number of annotated pages and query type.**

penalty as the number of annotated pages increases. Slope values are nearly uniform ($a \approx -0.024$), suggesting a similar drop in retrieval accuracy across most query types. The open-ended and enumerative types are the two exceptions: despite having the lowest NDCG@10 for low page counts, they also have the shallowest slope, which suggests that retrieval success on these queries is constrained by the model's fundamental difficulty in synthesizing multiple relevant sources rather than the volume of relevant context.

Performance by Query Generation Source Table 14 reports NDCG@10 of ColEmbed-3B-v2 stratified by query type and generation source. Queries generated with full page image access (Human+Image) consistently yield the highest scores, as annotators can anchor queries directly to visible content. Synthetic and human blind-contextual queries (SDG+Summary and Human+Summary) score similarly, with a mean difference of 0.044

Model	Size (B)	C.S.	Nucl.	Fin.	Phar.	H.R.	Ind.	Tele.	Average
Textual Retrievers									
Jina-v4	3	67.3	**48.2**	**56.5**	59.0	**58.8**	45.8	61.0	**56.7**
Qwen3-8B★	8	**73.5**	42.2	54.8	62.4	52.3	45.3	**66.0**	56.6
LFM2-350M	0.35	70.6	45.4	48.3	62.1	53.2	**47.9**	63.8	55.9
Mxbai Edge 32M	0.03	68.0	44.4	48.2	**62.5**	52.7	47.1	61.9	55.0
BM25S	-	64.7	45.9	49.9	56.9	49.6	45.6	58.3	53.0
Qwen3-0.6B★	0.6	70.5	39.7	51.5	57.4	46.2	42.4	59.7	52.3
GTE-ModernColBERT	0.15	63.6	41.7	39.8	62.0	46.2	44.6	59.7	51.1
BGE-M3★	0.57	63.6	34.3	43.9	54.7	45.3	39.0	54.3	47.9
Visual Retrievers									
ColEmbed-3B-v2	3	**78.6**	52.9	69.1	**67.6**	**65.4**	56.8	**71.7**	**66.0**
ColEmbed-3B	3	77.8	**53.4**	**69.5**	66.9	64.9	**57.0**	69.4	65.6
ColEmbed-1B	1	75.5	52.2	67.0	66.2	64.5	56.1	68.7	64.3
Jina-v4	3	74.2	52.4	66.1	65.2	64.6	55.9	68.7	63.9
ColNomic-7B	7	78.2	48.2	63.1	64.6	62.9	54.2	69.6	63.0
ColNomic-3B	3	75.5	45.5	63.0	63.7	62.6	52.8	68.6	61.7
ColQwen2.5	3	75.2	42.9	61.2	60.9	59.2	49.4	65.3	59.2
Nomic-7B★	7	70.9	42.3	57.6	63.8	55.9	48.5	62.0	57.3
ColQwen2	2	73.5	44.1	50.9	58.1	54.7	49.8	63.2	56.3
ColPali	7	72.5	38.1	43.3	57.7	53.3	47.0	59.2	53.0
Nomic-3B★	3	62.1	37.2	53.3	59.2	51.9	41.1	57.2	51.7
ColModernVBERT	0.25	59.7	42.0	50.4	56.6	47.0	43.9	55.2	50.7
ColSmol-256M	0.25	57.4	36.5	47.7	51.4	46.0	38.5	47.5	46.4

Table 9: **English-only retrieval performance (NDCG@10).** ★: single-vector models. Results are computed on the English queries of the English datasets.

Model	Size (B)	Phys.	Ener.	Fin.	Average
Textual Retrievers					
Jina-v4	3	44.0	**63.4**	**44.8**	**50.7**
Qwen3-8B★	8	**45.8**	60.2	37.6	47.9
Qwen3-0.6B★	0.6	43.8	54.9	28.5	42.4
BGE-M3★	0.57	38.3	53.1	28.4	39.9
BM25S	-	39.8	57.4	35.9	44.4
Visual Retrievers					
ColEmbed-3B-v2	3	48.2	67.5	48.2	**54.6**
ColNomic-7b	7	**48.5**	67.0	47.9	54.5
ColNomic-3b	3	**48.5**	**67.9**	46.8	54.4
Jina-v4	3	46.8	66.7	48.6	54.0
ColEmbed-3B	3	46.6	66.3	**48.9**	53.9
ColEmbed-1B	1	44.7	64.6	47.8	52.4
ColQwen2.5	3	47.8	62.3	43.6	51.2
Nomic-7B★	7	45.6	61.6	41.3	49.5
ColQwen2	3	43.9	55.6	26.5	42.0
Nomic-3B★	3	43.6	56.4	34.4	44.8
ColPali	7	43.2	50.5	23.6	39.1

Table 10: **French-only retrieval performance (NDCG@10).** ★: single-vector models. Results are computed on the French queries of the French datasets.

NDCG@10, confirming that the synthetic pipeline does not introduce a simplicity bias.

Performance by content type NDCG@10 by content type in Table 15 show that retrieval is more challenging for visual content, with Image perform-

Modality	English Datasets							French Datasets			
	C.S.	Nucl.	Fin.	Phar.	H.R.	Ind.	Tel.	Phys.	Ener.	Fin.	**Avg.**
Visual	39.4	25.5	28.4	27.5	30.0	21.4	31.4	**26.6**	25.2	22.9	27.8
Textual	35.4	23.1	24.5	24.2	27.4	16.5	29.0	25.8	23.9	20.4	25.0
Hybrid	**43.0**	**27.7**	**30.9**	**29.7**	**32.6**	**22.2**	**35.5**	26.5	**29.8**	**24.3**	**30.2**
Avg. # Pages for hybrid	*6.96*	*7.38*	*7.77*	*7.40*	*7.29*	*7.77*	*7.09*	*7.26*	*6.97*	*7.61*	*7.35*

Table 11: **Performance comparison of retrieval modalities (F1@10) on Jina-v4.** Evaluation is performed using the multimodal retriever Jina-v4. The Hybrid method combines the top-5 visual and top-5 textual matches, subsequently removing duplicates. The final row reports the average number of unique pages remaining in the hybrid set. The Hybrid setup constantly outperforms both textual and visual retrieval.

Query Language	NDCG@10
English	60.8
French	59.8
Portuguese	59.6
Spanish	59.6
Italian	59.1
German	57.9

Table 12: **ColEmbed-3B-v2 NDCG@10 by query language.**

Query Type	Slope a	Intercept b	R^2
Boolean	-0.0239	0.797	0.101
Numerical	-0.0255	0.742	0.059
Extractive	-0.0230	0.745	0.084
Compare-contrast	-0.0247	0.710	0.117
Enumerative	-0.0172	0.669	0.080
Multi-hop	-0.0237	0.680	0.114
Open-ended	-0.0129	0.577	0.057

Table 13: **Linear regression analysis of NDCG@10 decay with number of annotated pages, by query type.** The slope a represents performance sensitivity to retrieval context size, while the intercept b represents intrinsic difficulty at minimum context size.

Query Type	SDG+ Sum.	Human+ Sum.	Human+ Img.
Boolean	0.615	0.692	0.801
Compare-Contrast	0.573	0.514	0.743
Open-Ended	0.447	0.459	0.657
Extractive	0.568	0.623	0.743
Enumerative	0.449	0.559	0.697
Numerical	0.592	0.659	0.732
Multi-Hop	0.398	0.378	0.688
Mean	**0.521**	**0.565**	**0.726**

Table 14: **ColEmbed-3B-v2 NDCG@10 by query type and generation source. SDG+Sum.:** synthetic blind-contextual; **Human+Sum:**: human blind-contextual; **Human+Img:**: human with full page image access.

ing 10pp below Text. However, content type and query type are correlated in our benchmark: for instance, tables appear in numerical queries 2.2 $\times$ more often than the baseline, while images are over-represented in open-ended queries (Figure 13). Since numerical queries are easier than open-ended ones, we test whether the effect of content type is a byproduct of query type confounding. We fit an additive model that predicts performance as the sum of independent query-type and content-type effects. Figure 14 shows the residuals which measure deviation from this baseline. We see that most residuals are below 5pp, indicating that the two factors combine additively without significant interaction.

Content type	NDCG@10	Content type count
Text	59.3	17244
Chart	56.3	2364
Infographic	55.2	2814
Table	53.9	6480
Other	50.8	492
Image	49.3	1140
Mixed	45.1	1164

Table 15: **ColEmbed-3B-v2 NDCG@10 by content type.** Content type is labeled on each annotated page based on the nature of the query-relevant content delimited by the bounding boxes. One page may be tagged with several content types if it contains multiple relevant sections of distinct nature. The Mixed type corresponds to annotations encompassing several content types.

G VLM Filtering Effect on Query Distribution

Table 16 reports the recall/precision trade-off that motivated the choice of Qwen2.5-VL-32B-Instruct as the pre-filtering model.

To assess whether VLM pre-filtering introduces distributional bias, we compare query type, query format, and page modality distributions before and after the filtering step.

Table 17 shows that filtering causes minimal shifts across query types (all $|\Delta| \leq 3\%$), with the exception of Open-Ended queries (-5.4%), which

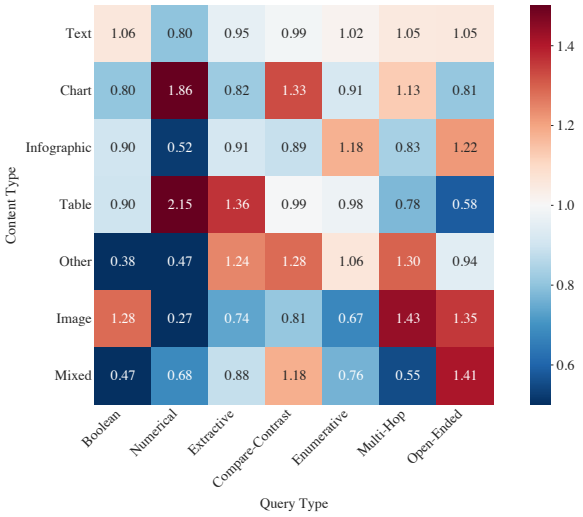


Figure 13: **Lift of query types by content type.** Each cell shows the ratio of observed query type frequency to baseline frequency for a given content type. Values >1 indicate over-representation (e.g., tables appear 2.15× more in numerical queries than expected), while values <1 indicate under-representation.

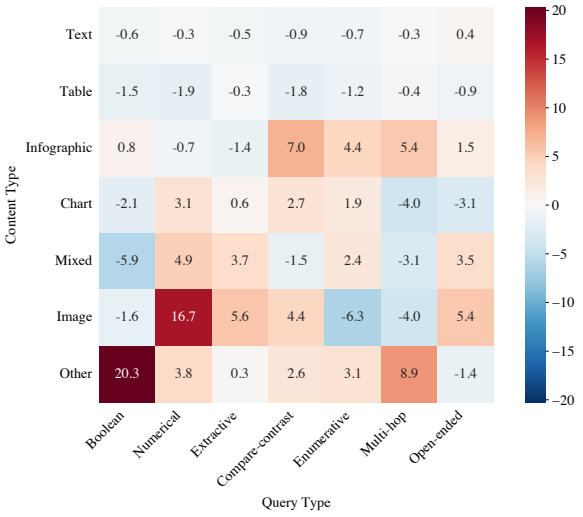


Figure 14: **Residuals from additive performance model.** Each cell shows the difference between observed NDCG@10 and the value predicted by an additive model of query type and content type main effects. Values near zero (white) indicate no interaction; positive values (red) indicate better-than-expected performance for that combination; negative values (blue) indicate worse-than-expected.

reflects the correct removal of overly vague queries. Table 18 shows similarly small shifts across query formats. Table 19 shows that the modality distribution of relevant pages is close to the source documents, confirming that complex visual content is not systematically excluded. The slight decrease in graphical elements use occurs because semantically irrelevant decorative items, such as logos, are naturally not targeted by queries.

Model	Recall	Precision
Qwen2.5-VL-32B	98%	32%
Gemini-2.5-Flash	78%	81%

Table 16: **VLM pre-filtering model selection.** Evaluated on 697 human-annotated query–page pairs from ViDoRe v2 (Macé et al., 2025) (103 positives, 594 negatives). We selected Qwen2.5-VL-32B as a high-recall sieve to avoid discarding ∼25% of relevant documents, leaving final relevance judgments to human annotators.

Query Type	Before (%)	After (%)	Δ
Boolean	13.88	16.07	+2.20
Compare-Contrast	13.73	13.25	−0.47
Enumerative	10.31	10.16	−0.15
Extractive	17.17	20.11	+2.94
Multi-Hop	8.27	8.06	−0.21
Numerical	6.32	7.42	+1.10
Open-Ended	30.33	24.93	−5.41

Table 17: **Query type distribution before and after VLM filtering.**

Query Format	Before (%)	After (%)	Δ
Instruction	31.32	30.50	−0.83
Keyword	20.00	17.51	−2.49
Question	48.67	51.99	+3.32

Table 18: **Query format distribution before and after VLM filtering.**

Source	Text	Table	Graphical
Relevant Pages	55.52%	21.20%	23.28%
Original Docs	54.10%	15.81%	30.09%

Table 19: **Modality distribution for relevant pages vs. original documents.** The proportion of text, table, and graphical elements is preserved after filtering, confirming that complex visual content is not systematically excluded.

H Bounding box annotations

Inter-annotator agreement Table 20 shows IoU and F1 scores between human annotations, to detail results of Section 3.3.

Validation on high-agreement pages To further characterize model grounding performance, we restrict evaluation to query–page pairs with high inter-annotator agreement (mean pairwise IoU $\geq$ 0.7).

	English Datasets							French Datasets			
Metric	C.S.	Nucl.	Fin.	Phar.	H.R.	Ind.	Tele.	Phys.	Ener.	Fin.	**Average**
IoU	0.500	0.476	0.462	0.615	0.474	0.502	0.526	0.443	0.470	0.503	**0.497**
F1	0.608	0.594	0.569	0.720	0.594	0.611	0.637	0.540	0.569	0.581	**0.602**

Table 20: **Inter-annotator agreement metrics on bounding box annotations.**

On this subset, Qwen3-VL-30B-A3B-Instruct and Gemini 3 Pro achieve both substantially higher than full-set scores, yet still well below human performance (21). This confirms that visual grounding is a genuine open challenge in current models.

Model	**Evaluation Set**	**F1**	**IoU**
Qwen3-VL-30B-A3B-Instruct	Full Set	0.089	—
	High-Agreement	**0.311**	**0.262**
Gemini 3 Pro	Full Set	0.065	—
	High-Agreement	**0.208**	**0.155**

Table 21: Model grounding performance comparison between the full evaluation set and the high-agreement subset (mean pairwise IoU ≥ 0.7).

Bounding box predictions Figure 27 shows the prompt used to generate final answers with inline bounding boxes for visual grounding, and Figure 15 reports bounding box localization F1 scores by dataset.

I Final answer evaluation

Evaluation setup Generated final answers are evaluated in a pass@1 setting using GPT 5.2 with medium reasoning effort as the LLM judge. The judge compares each generated answer against the ground-truth annotation and returns a binary correctness label. The answer generation and judge prompts are shown in Figure 25 and Figure 24 respectively. We evaluated Gemini 3 Pro with low thinking effort, GPT-5 with medium reasoning effort, as well as the thinking version of Qwen3-VL-235B-A22B.

To assess the reliability of our judge, we conducted 5 independent evaluation runs on a fixed set of Gemini 3 Pro outputs. Individual run scores showed minimal fluctuation (mean 72.09 %, $\sigma = 0.22$ %) and high internal consistency (Krippendorff's $\alpha = 0.91$), confirming that the judge is consistent given a fixed context.

End-to-End Pipeline Stability While the judge demonstrates high consistency on fixed inputs, the full evaluation pipeline introduces a second layer of variability: the model's generation process. To quantify the end-to-end variance under rigorous conditions, we performed 5 independent runs. For computational efficiency, we restricted this stress test to the most challenging corpus in each language: *Industrial Maintenance* (English) and *Finance* (French).

We measured an average score of 65.74 % with a standard deviation of 0.94 %. Crucially, the evaluation signal remains robust against generative noise, achieving a Krippendorff's α of 0.80. This agreement confirms that the end-to-end results are statistically reliable even when subjected to the most difficult evaluation scenarios.

Easy/hard query filtering To classify queries by difficulty, we prompt a panel of 6 LLMs to answer each query without access to any corpus context. We select GPT-5-nano, GPT-5-mini, GPT-5, Qwen3-VL-30B-A3B, Gemini 2.5 Flash, and Gemini 2.5 Pro to span different model families and capability levels. Each model receives only the query text and is asked to provide a direct answer with the prompt in Figure 23. Answers are evaluated for correctness using the same GPT-5.2 judge described above. A query is labeled *easy* if at least one model answers correctly, and *hard* otherwise. Table 22 reports per-model accuracy and the resulting proportion of easy queries for each dataset. The distribution varies substantially across domains: knowledge-intensive datasets such as Computer Science and Physics have over 85% easy queries, while domain-specific datasets such as Finance and Energy contain fewer than 35% easy queries, reflecting the specialized nature of their content.

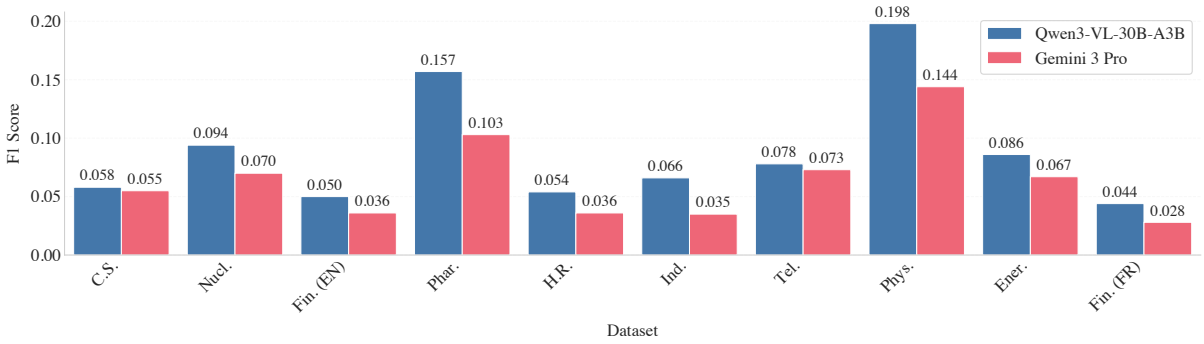


Figure 15: **Model bounding box localization performance**. Each F1 score measures the zone-based overlap between model-generated bounding boxes and human annotations, using the annotator yielding the highest F1.

	English Datasets					French Datasets			
Model	C.S.	Fin.	Phar.	H.R.	Ind.	Phys.	Ener.	Fin.	Total
GPT-5-nano	74.4	7.4	30.5	12.9	15.6	74.2	14.0	9.1	29.8
GPT-5-mini	79.1	13.3	37.4	17.6	20.5	80.1	13.6	13.1	34.3
GPT-5	76.3	25.2	50.8	29.9	32.2	80.5	26.0	22.2	42.9
Qwen3-VL-30B-A3B	60.9	3.9	19.2	6.3	9.9	60.9	6.8	4.4	21.5
Gemini 2.5 Flash	66.1	8.7	30.8	13.5	15.9	63.6	14.9	13.1	28.3
Gemini 2.5 Pro	70.2	16.8	29.4	15.4	23.7	63.9	20.5	20.0	32.9
Easy queries (%)	86.5	31.7	57.1	36.5	38.9	86.4	32.5	30.0	48.6

Table 22: **Percentage of queries correctly answered by LLMs without corpus context.** A panel of 6 LLMs is asked to answer the queries of the 8 public datasets without access to any corpus context. Queries correctly answered by at least one of the 6 models are classified as *easy* queries, while the rest are labeled as *hard*. Easy queries account for 48.6 % of all the queries.

J Visual grounding examples

Qualitative analysis reveals distinct failure modes. Gemini frequently produces off-by-one page indexing errors: the predicted coordinates would correctly localize the target content if applied to an adjacent page. The two models also differ in box granularity: Gemini tends to draw tight boxes around individual elements (e.g., a single table cell or text line), whereas Qwen3-VL generates larger boxes encompassing entire sections or paragraphs, more closely matching human annotation patterns. Figures 16 and 17 illustrate these tendencies across four dataset pages: Qwen3-VL's bounding boxes are comparatively wide and encompass entire page elements (pages (a), (c), and (d)), while Gemini 3 Pro's visual grounding is more precise (pages (b) and (c)). This difference in granularity partially explains Qwen3-VL's higher F1 scores, as broader boxes are more likely to overlap with the ground-truth zones used in our evaluation. Both models exhibit errors and omissions: in page (b), the chart is not labeled by Qwen3-VL, and in page (d), Gemini 3 Pro predicts incorrect bounding boxes for the bottom table while Qwen3-VL provides grounding for the wrong table.

K Instructions given to Annotators

Query Generation Figure 18 details step-by-step instructions to annotators to generate queries from summaries and images.

Query-Page Relevancy linking Figure 19 details the step-by-step instructions provided to annotators for assessing page relevance, identifying content modalities, and localizing evidence via bounding boxes. Table 23 gives the definitions of relevancy scores used by the human annotators.

Score	Label	Definition
2	Fully Relevant	The page contains the complete answer.
1	Critically Relevant	The page contains facts or information required to answer the query, though additional information is required.
0	Not Relevant	Provides no information relevant to the query.

Table 23: **Relevance definitions used for page-level annotations.**

L Prompts

All the prompts used for both dataset generation and evaluations are detailed from Figure 20 to Figure 27.

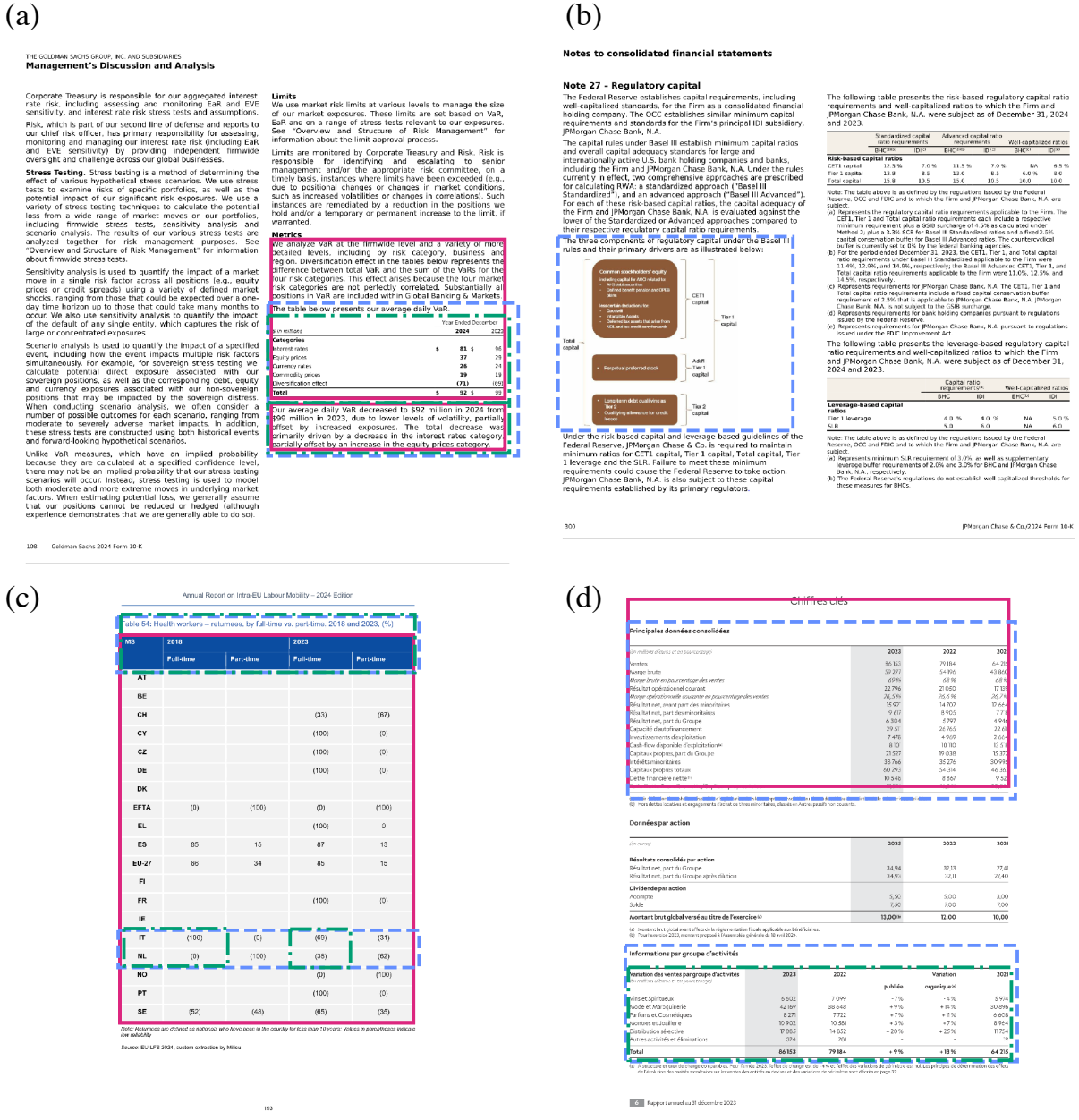


Figure 16: **Visual grounding comparative examples for Qwen3-VL-30B-A3B.** Each panel shows a document page with Qwen3-VL's predicted bounding boxes (solid magenta) and human bounding boxes (dashed blue and green, one color per annotator). Corresponding datasets and queries: (a) finance_en: What was the average daily Value at Risk (VaR) for Goldman Sachs during 2024?, (b) finance_en: List the 3 components of regulatory capital under Basel III, and determine the role of each component., (c) hr_en: Analyze how full-time employment among returning health workers evolved in the Netherlands and Italy from 2018 to 2023, and describe the differences in their employment trends., (d) finance_fr: Croissance Mode Maroquinerie vs Vins Spiritueux 2023 performance

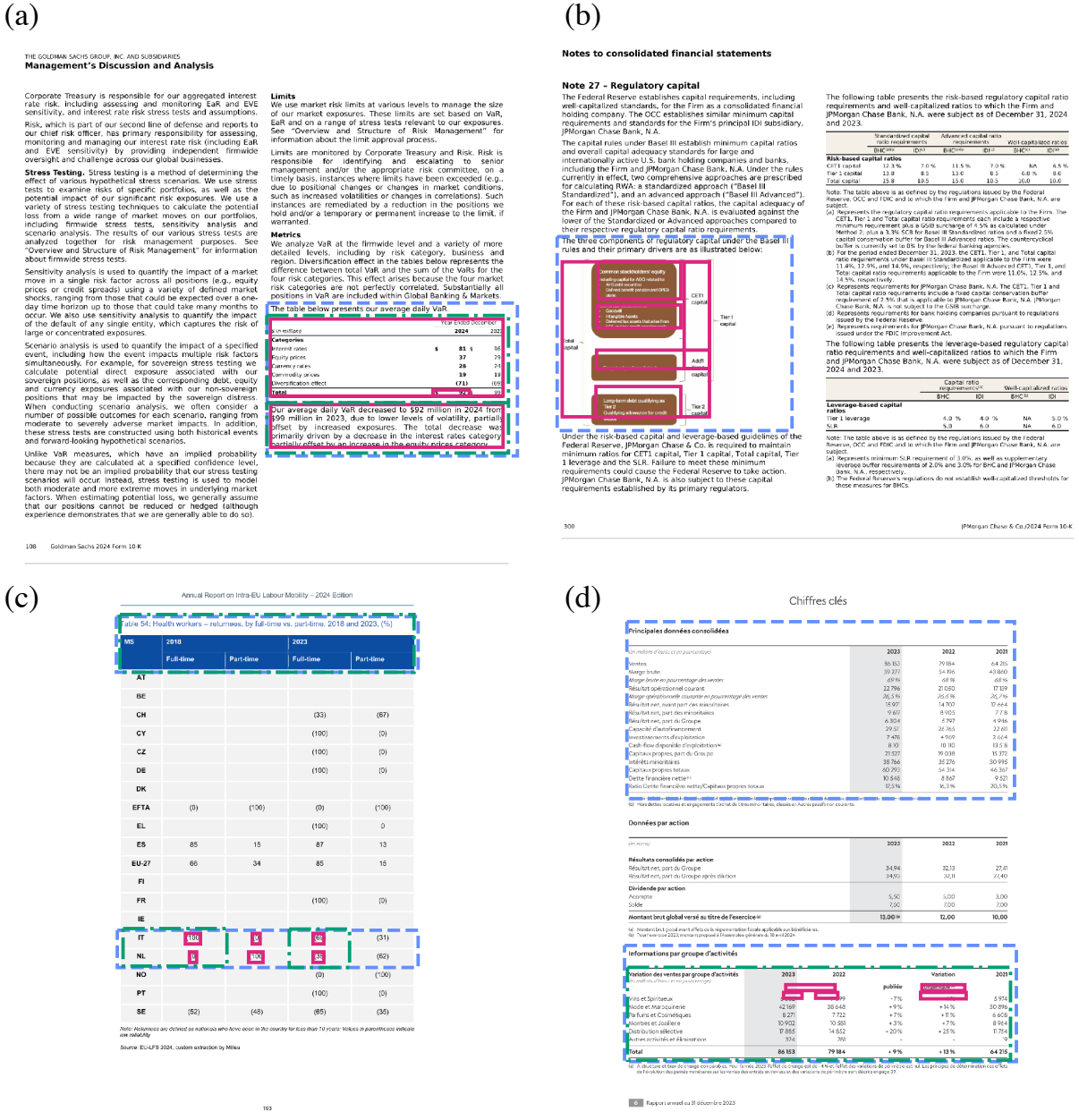


Figure 17: **Visual grounding comparative examples for Gemini 3 Pro.** Each panel shows a document page with Gemini's predicted bounding boxes (solid magenta) and human bounding boxes (dashed blue and green, one color per annotator). Corresponding datasets and queries: (a) finance_en: What was the average daily Value at Risk (VaR) for Goldman Sachs during 2024?, (b) finance_en: List the 3 components of regulatory capital under Basel III, and determine the role of each component., (c) hr_en: Analyze how full-time employment among returning health workers evolved in the Netherlands and Italy from 2018 to 2023, and describe the differences in their employment trends., (d) finance_fr: Croissance Mode Maroquinerie vs Vins Spiritueux 2023 performance

```
# Step 1
   The annotator will be provided with content(text summary or image(s)) and a list of instructions
    on the queries that are expected.

# Step 2
    Read and analyse the content.

# Step 3
   - The annotator writes a series of queries that can, supposedly, be answered by the summarized
    content. It is okay if the information needed to answer the question is in other parts of the
    document or not explicitly written in the summary.
   - If the summary is not adapted to a specific type or format of questions, the annotator may skip.
    - They should follow the number of queries asked for each category.
    - They should follow the expected type of queries provided and the format.
```

Figure 18: **Instructions given to human annotators to create queries**

```
# Task overview
In this task, the annotator will be provided with a query and pages that are supposed to be relevant
to answer the query. The annotator's goal is to rate the relevance in answerability of each page with
respect to the query.

# Step 1:
    - Review the query and pages to get an understanding of the content and domain

# Step 2:
    Rate the query quality.
    - If adheres to guidelines > Good (1)
    - If doesn't adhere to guidelines> Poor(0)
    - If the query is Poor quality, skip the task.

# Step 3:
    - For each page, rate the relevance with respect to the query
    - If page completely answer query > Fully Relevant(2)
    - If page contains information required to answer the query > Critically Relevant(1)
    - If page contains no relevant information > Not Relevant(0)

# Step 4:
   - For each page, annotate the modalities in which the relevant information is located concerning
   the query-page link, relevant information can be located in multiple modalities at the same time:
Modality : ["text", "table", "chart", infographic", "image", "other"]
    - If relevance score = 0, modality may be "N/A"

# Step 5:
   - For each page, draw bounding boxes around the relevant text/chart/image/infographic (if any)
    - If relevance score = 0, do not draw a bounding box

# Step 6:
    - Repeat steps 3, 4 and 5 for all pages

# Step 7:
    - Propose an answer to the query, given the relevant pages.
    - If the query is not answerable, rate it as "unanswerable"
```

Figure 19: **Instructions given to human annotators to annotate query-page relevancy**

```
<mission>
You are an assistant specialized in visual document understanding tasks. You will be given a context,
summarizing the content of a section or multiple document sections. Your goal is to carefully analyze
the context and to solve a series of tasks related to its content. You are tasked with generating
query-answer pairs. Your queries will be used to simulate a user unfamiliar with the specific content
of the page, and who is looking for information in a large knowledge base through a search engine.
The user does not have access to the document and is looking for information that can be present in
any document in the knowledge base.
</mission>

<definitions>
- A query is said to be fully answerable if the page contains a precise and complete answer to the
query.
- A query is said to be partially answerable if the page contains relevant information that is
directly related to the query but some key information is missing and must be retrieved in other
pages or documents in order to give a precise and complete answer.
- An open-ended query is an explanatory or descriptive query that synthesizes  information; may be
broad in scope and focused on qualitative aspects of the summary
- A compare-contrast query is a query that requires comparing and/or contrasting multiple entities or
topics that are closely related to each other
- An enumerative query is a query that asks to list all examples that possess a common property,
optionally requesting details about the specifics of each example.
- A numerical query is a query that asks for a specific number or calculated number given a summary.
The query should require more than simply reading numbers directly from the page.
- A boolean query is a yes/no query that may involve multiple steps of reasoning.
- An extractive query is a clear and specific query that can be answered using only a specific piece
of information.
- A multi-hop query is a complex query that requires retrieving and integrating information from
multiple sources or steps to produce a complete answer.
- A question query is a complete sentence that ends with a question mark, typically used to seek
specific information or clarification.
- A keyword query is a brief, often fragmented phrase or set of terms used to search or filter
information, without forming a full grammatical sentence.
- An instruction query is a directive that describes a task to be performed on the documents, often in
the form of a command or request.
</definitions>
<rules>
<queries>
- Generate queries only in {{ language }}.
- Make queries diverse, natural, and plausible for someone unfamiliar with the document.
- Each query must be standalone; do not reference “the page”, “the table”, “the figure”,
“the document”, “the text”, “the table of contents”, etc.
- Rephrase; avoid copying wording from the source so semantic matching, not surface matching,
is tested.
- You may include queries about relationships or trends often shown in tables/figures/graphs,
but never refer to a specific table/figure.
- Avoid overly generic queries that apply to any document.
- Keep each query concise ({{ length }} words).
- When appropriate, write multi-hop queries that integrate information across the provided pages.
</queries>
</rules>
<instructions>

Used Documents: {{ document_names }}
<summary>
{{summary}}
</summary>
Using the provided context, generate a {{ difficulty }}, {{ reasoning_type }}, {{ answerability }}
query. The query should be {{ query_type }} using the provided context and have the format of a
{{ query_format }} query.
The query should be self-sufficient and related to the context.
</instructions>
```

Figure 20: **Prompt used generate synthetic queries with the NeMo Data Designer tool**

```
<mission>
You are an assistant specialized in visual document understanding tasks. You will be given a document
page by page and a question. Your goal is to carefully analyze the page and say if it is related to
the question's answer. You are tasked with generating question-page affiliation as well as the
question answer if it exists in the page.
</mission>

<definitions>
- A question is said to be fully answerable if the corresponding page contains a precise and complete
answer to the question.
- A question is said to be partially answerable if the corresponding page content is necessary to
answer the question but some key information is missing.
- A question is said to be unanswerable if the corresponding page contains information related to the
question's topic or domain but upon closer inspection does not contain information that is useful to
answer the question. Or if the page has no link whatsoever with the query.
</definitions>

<rules>
<page_affiliation>
- Be sure to put the relevance (and only that) between the tags <relevance>...</relevance>.
The possible values are:
    <relevance>fully answerable</relevance>,
    <relevance>partially answerable</relevance>
    <relevance>unanswerable</relevance>.
- Be very careful when doing your page affiliation. Only say a page is relevant when it really is.
</page_affiliation>
<answers>
- You must generate the answer between the tags <answer>...</answer>. Between these tags, you should
only put the answer to the question.
- You must generate answers in the following language: {language}
- Your answers should be complete sentences.
- When the question is ambiguous, your answer should state that there is an ambiguity in the question.
- You should always generate the answer based on all the information available on the page, even if
the question was generated only on part of the page.
</answers>
</rules>

<instructions>
Return if the following question is "fully answerable" "partially answerable" or "unanswerable" based
on the content of the page between the tags <relevance> ... </relevance>.
If the question is answerable, provide the answer.

Here is the question :
{{ query }}

And there is the page content :

</instructions>
```

Figure 21: **Prompt used to pre-filter the irrelevant pages for a given query**

```
You are given a set of document pages (images), a query, and a list of one or more
proposed answers.

Query :

{{ query }}

Proposed Answers:

{{ answers }}

Your task is to carefully analyze the provided pages, the query, and the proposed answers. You must
return a single, syntactically correct JSON object with the following structure:
```json
{
    "reasoning": "<string>",
    "information_in_pages": <true or false>,
    "answer_correctness": [<true or false>, ...],
    "reformulated_answer": "<string>"
}
```
Instructions for each field:
- reasoning: Explain the logic for each boolean in the `answer_correctness` list.
For each proposed answer, state why it is correct or incorrect, citing specific evidence from the
document pages.

- information_in_pages: Set to `true` if the information needed to definitively answer the query
is present in the pages. Otherwise, set to `false`.

- answer_correctness: A list of booleans, corresponding to each proposed answer in the original
order. Use `true` if the answer is verifiably correct based on the pages and `false` otherwise.

- reformulated_answer: A single string containing the most precise and correct answer to the query,
derived only from information in the pages. If any of the proposed answers are correct, use them as
a basis for synthesizing this improved answer. The reformulation must be concise and factual.

Important rules:
- Base your entire analysis strictly on the content of the provided document pages.
Do not use outside knowledge.

- Do not invent, infer, or assume information that is not explicitly stated in the pages.

- Always provide a string for the `reformulated_answer`, even if no correct answer can be formed
from the text.

- Your final output must be only the JSON object.
```

Figure 22: **Prompt used to merge human annotators answers**

```
Give a very precise and concise answer to the following query.
If you are unable to answer, output 'I don't know'.
Query: {{ query }}
```

Figure 23: **Easy/hard query filtering prompt**

```
You are an expert judge evaluating the accuracy of a test answer against a gold-standard true answer.
Your goal is to determine if the test answer captures the essential "core information."

### Evaluation Criteria:
- Correct: The test answer contains all core information of the true answer. Minor omissions of
non-essential details or the addition of minor, non-contradictory information should still be marked
as "Correct."
- Partially Correct: The test answer captures some of the core information, but suffers from
significant omissions or includes substantial extra information that was not requested or present in
the true answer.
- Incorrect: The test answer is fundamentally wrong, contradicts the true answer, or misses the core
information entirely.

### Input Data:
Query: {{ query }}

True Answer: {{ true_answer} }

Test Answer: {{ test_answer }}

### Output Format:
Provide a very brief explanation for your judgment. You must output your final response in a JSON
format with two fields: "explanation" and "judgment" (which must be "Correct", "Partially Correct",
or "Incorrect").
```

Figure 24: **Judge prompt used for end to end evaluation**

```
You are an expert at answering query based on documents.

Here is a list of relevant documents: {{ documents }}

Based on the above documents, answer the following query: {{ query }}

Keep the response short when appropriate. Output the answer only.
```

Figure 25: **Answer generation prompt used for end to end evaluation**

```
English text to translate: {{ query }}

Translate the English text above to French.
Make sure you follow the format of the English text. Don't change acronyms.
Follow the following json schema.
{
    "french_translation": ...
}
```

Figure 26: **Query translation from English to French prompt**

```
# Role and Objective
- Serve as an expert in document analysis and visual grounding.
- Given a query and multiple document page images, provide a natural language answer with inline
grounding references.

# Instructions
- Analyze all provided pages to answer the query comprehensively.
- For each piece of information used in your answer, provide visual grounding by including bounding
box coordinates of all the sections of the document that help answer the query.
- Use this format to include the list of all bounding boxes of image N:
<bboxes image="N">[[x_{min}, y_{min}, x_{max}, y_{max}], ...]</bboxes>
  - image="N" specifies the 0-indexed page number (0=first page, 1=second page, etc.)
  - Include bounding boxes inline in your answer, immediately after mentioning the relevant
  information
  - A given page may contain multiple non-contiguous sections that help answer the query. In this
  case, you must output the list of the bounding boxes of all these sections.
  - You must group all the bounding boxes of a given page into a single
    <bboxes image="N">...</bboxes> tag.

# Grounding Principles
- DO NOT output more than 5 bounding boxes per page.
- Adjacent logical units must be enclosed in a single, continuous bounding box.
- Return multiple bounding boxes only if information is clearly independent and separated by
significant non-relevant content.

# Output Format
- Provide a natural language answer to the query.
- Embed grounding tags directly inline where relevant information is discussed.
- Example: "The valuation technique described on page 1
  <bboxes image="0">[[120, 450, 890, 670], [100, 800, 330, 960]]</bboxes>
  uses discounted cash flow analysis."
```

Figure 27: **Bounding box prediction prompt**