

# Exploration of Augmentation Strategies in Multi-modal Retrieval-Augmented Generation for the Biomedical Domain\*

\*A Case Study Evaluating Question Answering in Glycobiology

Anonymous authors

**Abstract**—Multi-modal retrieval-augmented generation (MM-RAG) promises grounded biomedical QA, but it is unclear when to (i) convert figures/tables into text versus (ii) use optical character recognition (OCR)-free visual retrieval that returns page images and leaves interpretation to the generator. We study this trade-off in glycobiology, a visually dense domain. We built a benchmark of 120 multiple-choice questions (MCQs) from 25 papers, stratified by retrieval difficulty (easy text, medium figures/tables, hard cross-evidence). We implemented four augmentations—None, Text RAG, Multi-modal conversion, and late-interaction visual retrieval (ColPali)—using Docling parsing and Qdrant indexing. We evaluated mid-size open-source and frontier proprietary models (e.g., Gemma-3-27B-IT, GPT-4o family). Additional testing used the GPT-5 family and multiple visual retrievers (ColPali/ColQwen/ColFlor). Accuracy with Agresti-Coull 95% CI: Gemma-3-27B-IT, Text and Multi-modal conversion outperformed OCR-free retrieval (0.722–0.740 vs. 0.510 average accuracy). With GPT-4o, Multi-modal achieved 0.808, with Text 0.782 and ColPali 0.745 close behind; within-model differences were small. In follow-on experiments with the GPT-5 family, the best results with ColPali and ColFlor improved by 2 percentage points to 0.828 in both cases. In general across the GPT-5 family, ColPali, ColQwen, and ColFlor were statistically indistinguishable; ColFlor matched ColPali while being far smaller. GPT-5-nano trailed larger GPT-5 variants by roughly 8–10%. Pipeline choice is capacity-dependent: converting visuals to text lowers the reader burden and is more reliable for mid-size models, whereas OCR-free visual retrieval becomes competitive under frontier models. Among retrievers, ColFlor offers parity with heavier options at a smaller footprint, making it an efficient default when strong generators are available.

## I. INTRODUCTION

The rapid proliferation of Large Language Models (LLMs) has transformed numerous domains, including biomedical question answering (QA). Advanced systems such as Med-PaLM 2 have demonstrated expert-level performance on standardized medical examinations [1], leveraging well-established biomedical benchmarks such as BioASQ [2] and PubMedQA [3]. Despite these successes, their reliability in specialized scientific domains remains limited due to training data constraints. This highlights the need for external knowledge

grounding to mitigate hallucinations and improve factual accuracy [4].

Retrieval-Augmented Generation (RAG) has emerged as a robust alternative or complement to fine-tuning (FT), dynamically providing relevant contextual information to LLMs during inference [5]. Compared with FT, RAG typically achieves stronger performance at substantially lower computational cost and does not require retraining when new knowledge becomes available [6]. This property makes RAG particularly suitable for rapidly evolving biomedical literature.

While text-based RAG systems are well established, scientific research often involves multi-modal content extending beyond text. Our study focuses on glycobiology—a visually dense and technically demanding domain encompassing complex molecular structures, pathway diagrams, and tabular datasets [7]. Prior work has shown that even advanced LLMs struggle with glycobiology-related queries, frequently producing inconsistent or fabricated responses [8]. This underscores the need for multi-modal RAG (MM-RAG) systems capable of processing heterogeneous data modalities.

In this preliminary study, we compare two primary MM-RAG paradigms. The first converts multiple modalities into text following a conventional pipeline: PDF documents are parsed into structured elements—via direct text extraction or Optical Character Recognition (OCR)—while figures and tables are transformed into textual descriptions using summarization techniques [9]. The textual content is subsequently chunked and embedded for retrieval via standard vector databases. This approach simplifies decoupling between retrieval and generation but risks information loss and pipeline complexity.

The second paradigm employs vision-based document retrieval to circumvent conversion-related limitations. These methods treat entire document pages as images and utilize Vision-Language Models (vision-language models (VLMs)) to generate embeddings directly. A notable example is ColPali [10], which adapts the late-interaction mechanism from ColBERT [11] to the visual domain, enabling fine-grained similarity matching between visual and textual embeddings. Related approaches such as VisRAG also leverage VLM-based retrievers and generators to preserve maximal information from original documents [12]. In such setups, the downstream

This work was supported by European Union under Horizon Europe [grant number 101159018]; by University of Maribor under the 2023 internal call Strengthening Researchers' Programme Cores, field of Data Science and Artificial Intelligence in Biomedicine.

LLM assumes greater responsibility for visual reasoning, effectively replacing traditional PDF parsing and summarization components.

This study investigates whether the performance of OCR-free, vision-based document retrieval depends on the multi-modal reasoning capacity and scale of the downstream generative model in the context of multiple-choice question (MCQ) answering from glycobiology literature. We evaluate three RAG strategies—standard text-based RAG, modality-converting MM-RAG, and vision-based ColPali—across both large proprietary models (e.g., the OpenAI GPT-4o family) and smaller open-source alternatives (e.g., Gemma-3-27B-IT). Furthermore, we assess different late-interaction visual retrievers—ColPali [10], ColFlor [13], and ColQwen [14]—using the recently released OpenAI GPT-5 family.

Our findings reveal a trade-off between pipeline simplicity and dependence on model reasoning capacity, establishing an initial baseline for developing reliable and trustworthy multi-modal RAG systems in specialized biomedical domains.

## II. RELATED WORK

The biomedical domain is inherently multi-modal, as clinicians and researchers routinely integrate information from medical imaging, laboratory results, electronic health records (EHRs), and genomics [15], [16]. This complexity makes it a natural setting for developing multi-modal artificial intelligence (AI) systems. Notable progress has been achieved in multi-modal question answering (MMQA), with models such as LLaVA-Med [17], Med-Flamingo [18], and Med-PaLM M [19], which jointly reason across textual and visual modalities in clinical contexts. However, public benchmarks have revealed potential data contamination in newer LLMs, motivating the use of private, domain-specific datasets for evaluation.

Biomedical AI represents a key application area for Retrieval-Augmented Generation (RAG), characterized by a vast and rapidly evolving knowledge base where factual accuracy and evidence-based reasoning are paramount [20], [21]. Standard LLMs are static and prone to hallucinations, making RAG critical for building reliable clinical systems [20]. A systematic review of 30 studies identified diagnostic support, EHR summarization, and medical QA as the most prevalent RAG applications [21]. A meta-analysis of 20 studies comparing baseline LLMs with RAG-augmented counterparts reported a pooled odds ratio of 1.35 (95% CI: 1.19–1.53), demonstrating significant performance gains through RAG integration [20].

For instance, a biomedical QA system employing a fine-tuned Mistral-7B model that retrieved information from PubMed and medical encyclopedias achieved a BERTScore F1 of 0.843 [22], [23]. Similarly, systems such as MEDGPT [24] have demonstrated the practical utility of RAG for diagnostics and automated report generation.

Despite these advances, several challenges persist. The technical terminology and structural density of biomedical

literature often introduce retrieval noise, resulting in sub-optimal generation [21]. Furthermore, biomedical content is fundamentally multi-modal: research articles, clinical guidelines, and EHRs frequently include visual components such as graphs, flowcharts, and tables that are integral to interpretation. Yet, most existing biomedical RAG implementations remain text-focused. The same systematic review confirming RAG’s clinical benefits revealed that only three of the twenty analyzed studies explicitly incorporated tabular or visual data—and typically by converting these modalities into text [20]. This limited engagement with native multi-modal content underscores the need for architectures capable of directly processing complex visual information.

To address this gap, the MRAG-Bench benchmark was introduced to evaluate vision-centric multi-modal retrieval-augmented models [25]. It contains multiple-choice questions requiring visual reasoning across scenarios involving perspective shifts, occlusion, and temporal or geometric transformations [25]. The benchmark highlights a substantial gap between machine and human visual reasoning capabilities: when provided with ground-truth visual knowledge, GPT-4o’s accuracy improved by only 5.82%, whereas human participants improved by 33.16% [25]. These findings quantitatively support the hypothesis that the performance of vision-based RAG pipelines depends strongly on the scale and multi-modal reasoning ability of the underlying generative model.

## III. MATERIALS AND METHODS

### A. Benchmark

A private benchmark dataset was constructed by two domain experts, comprising 120 multiple-choice questions (MCQs) derived from 25 original research and review manuscripts. Each question contained four possible answers and the corpus spanned core glycobiology concepts and applications, ranging from population-scale IgG glycomics, immune and inflammatory regulation, and endocrine/aging effects, to cardio-metabolic, gastrointestinal, pulmonary, and oncologic disease phenotypes (Appendix A). Questions were also categorized by retrieval difficulty by consensus by the two domain experts: *easy* when the answer appeared directly in the text, *medium* when it was presented in tables or figures, and *hard* when it required integrating information across text, figures, supplementary tables, or cited references. Iterative manual refinements were performed through review of model-generated explanations, during which five items were reclassified to ensure consistent difficulty labeling.

### B. Vision-Language Model Selection

The selection of open-source vision-language models (VLMs) was constrained by local inference resources—specifically a single NVIDIA H100 80GB PCIe GPU. Considering activation memory and framework overhead, models up to approximately 30 billion parameters were feasible under 16-bit precision. We focused on models adapted from general multi-modal LLMs to biomedical applications [26], employing a generate-then-filter pipeline to

synthesize diverse visual-instruction data from biomedical image-caption pairs. The evaluated models included Qwen2-VL-2B-Instruct [27], LLaVA-NeXT-Llama3-8B [28], and Llama-3.2-11B-Vision-Instruct [29]. We additionally tested Google’s Gemma-3-27B-IT [30]. They were deployed using vLLM [31] via docker (example in Appendix C).

For proprietary baselines, the OpenAI GPT-4o family (gpt-4o, version 2024-11-20; gpt-4o-mini, version 2024-07-18) and GPT-5 family (gpt-5, gpt-5-mini, gpt-5-nano; all version 2025-08-07) were accessed via API under a GDPR-compliant Data Processing Addendum (DPA) [32].

### C. Multi-Modal RAG Framework

The multi-modal RAG framework consisted of three modular components: a document parser, an open-source vector store, and a vision-language model for text-visual alignment at inference. IBM’s *Docling* served as the document parser [33], and *Qdrant* as the vector database [34], supporting both single- and multi-vector representations. Four vector store configurations were developed: (i) *text-only conversion*, in which all modalities were summarized into text; and (ii) *visual late-interaction* variants (ColPali), in which document pages were represented as image embeddings. Retrieval employed a late-interaction mechanism [11] for fine-grained similarity matching. Standard semantic embeddings (BAAI/bge-base-en-v1.5) [35] were used for the text-only and multi-modal-text configurations. A general prompt template was used for creating table and figure summaries (Appendix B). *Docling*, *Qdrant* and the open-source models were deployed on-premise on a GPU server (NVIDIA H100 80GB PCIe GPU) using docker and accessed through api (example configurations are in Appendix C).

We compared multiple augmentation strategies supplied to the LLM: *None* (query only), *Text* (nearest text chunks via standard RAG), *Multi-Modal* (raw figures and tables with corresponding summaries), and vision-based retrievals (*ColPali*, *ColQwen*, *ColFlor*), in which the most similar document pages were retrieved and passed directly to the LLM.

### D. Vision-Based Retrievers

We selected vision-based retrievers that demonstrated strong performance on the ViDoRe v2 benchmark [36]—particularly on healthcare-related subsets—and that were parameter-efficient, such as ColFlor. Specifically, we used the repositories `vidore/colpali-v1.3-merged` for ColPali, `vidore/colqwen2-v0.2` for ColQwen, and `ahmed-masry/ColFlor` for ColFlor.

**ColPali** (`vidore/colpali-v1.3-merged`) is a vision-based document retrieval model built on the PaliGemma-3B vision-language backbone. It extends SigLIP by feeding patch embeddings into PaliGemma to produce ColBERT-style multi-vector representations of pages [37]. The model (2.92 billion parameters) encodes entire page images—including text, layout, figures, and tables—into patch embeddings, then matches queries to pages using late-interaction scoring (e.g., MaxSim) over token-patch

pairs [38]. Strengths include preservation of visual layout and non-textual information, elimination of OCR and layout parsing, and strong retrieval accuracy on visual document benchmarks. Limitations include high memory and storage requirements due to multi-vector embeddings and increased computational cost for large document collections.

**ColQwen2** (`vidore/colqwen2-v0.2`) follows the ColPali paradigm but leverages the Qwen2-VL-2B backbone. It adopts a multi-vector retrieval scheme and incorporates adapters on top of Qwen2-VL-2B [39]. The Qwen2 backbone provides vision-language alignment across image and text modalities. Owing to its modular, adapter-based architecture, ColQwen2 facilitates efficient model switching and reduced update cost. However, its performance depends on the quality of adapter tuning; suboptimal alignment can degrade retrieval quality relative to ColPali. It also inherits the same storage and matching complexity constraints associated with multi-vector retrieval.

**ColFlor** (`ahmed-masry/ColFlor`) is a lightweight visual retriever optimized for footprint and speed. According to its model card, ColFlor contains approximately 174 million parameters—roughly  $17\times$  smaller than ColPali [40]. It achieves query encoding  $\sim 9.8\times$  faster and image encoding  $\sim 5.25\times$  faster than ColPali, with only a  $\sim 1.8\%$  performance drop on text-rich English documents [41]. Architecturally, ColFlor employs Florence-2’s DaViT vision encoder followed by a BART text encoder to produce contextualized visual embeddings, projected into a 128-dimensional latent space for retrieval via late interaction [41]. Advantages include low computational requirements, high throughput, and suitability for resource-constrained environments, whereas limitations include reduced performance on highly visual or non-English documents and limited capacity for capturing fine-grained visual detail due to its smaller backbone.

### E. Evaluation

Performance was quantified using mean accuracy, defined as the proportion of correctly answered MCQs. With four equally likely options, random guessing yields an expected accuracy of 0.25. We evaluated both overall and difficulty-stratified performance. Two primary experiments were conducted with a general prompt template used (Appendix D).

First, we compared proprietary GPT-4o models with the open-source Gemma-3-27B-IT across all augmentation strategies, performing ten runs per configuration—five with permuted answer orders and five without—to assess potential benchmark contamination. Second, we compared vision-based retrievers (ColPali, ColQwen, ColFlor) using the GPT-5 family, with five evaluation runs each.

All accuracy estimates were reported with Agresti-Coull 95% confidence intervals, recommended for robust estimation of binomial proportions [42]. Pairwise comparisons between augmentation strategies applied Bonferroni correction for multiple testing.

TABLE I  
INITIAL MODEL EVALUATION WITH REGARD TO THE PROPOSED  
FRAMEWORK.

Model	Augmentation			
	None	Text	Multi-modal	ColPali*
AdaptLLM/biomed	0.258	0.192	0.283	0.258
-LLaVA-NeXT-Llama3-8B				
AdaptLLM/biomed	0.217	0.283	0.283	/
-Qwen2-VL-2B-Instruct				
AdaptLLM/biomed	0.383	0.508	0.542	0.475
-Llama-3.2-11B-Vision-Instruct				
google	<u>0.483</u>	<u>0.692</u>	<u>0.767</u>	<u>0.558</u>
/gemma-3-27b-it				
gpt-4o-mini	0.458	0.708	0.675	0.558
<b>gpt-4o</b>	<b>0.592</b>	<b>0.792</b>	<b>0.833</b>	<b>0.758</b>

\*AdaptLLM/biomed-Qwen2-VL-2B-Instruct ColPali augmentation did not produce relevant answers

#### IV. RESULTS

From the initial testing (Table I), smaller open-source models—even when fine-tuned—struggled on domain-specific MCQs. Augmentation improved accuracy by roughly 12–26%. Among open-source models, gemma-3-27b-it achieved the strongest performance and was selected for detailed evaluation. ColPali augmentation performed well primarily with gpt-4o.

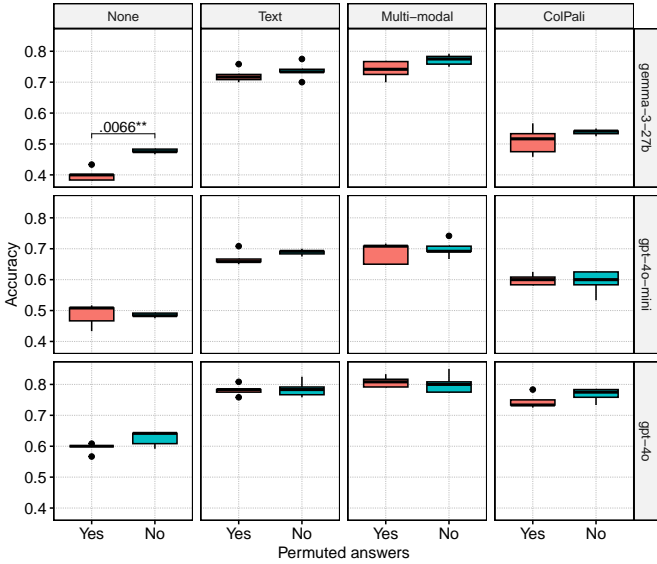


Fig. 1. Boxplot and significant  $p$ -values for accuracy across selected models and augmentations.

Because contamination in public benchmarks can undermine evaluation robustness, we tested whether answer-order permutations affected performance. As expected (Figure 1), we found no statistically significant differences at  $\alpha = 0.05$ . The closest case was gemma-3-27b-it with no augmentation (unadjusted  $p = 0.0066$ ), consistent with its lower accuracy

and higher variability (0.477; 95% CI [0.437, 0.517]) versus 0.400 (95% CI [0.363, 0.440]).

TABLE II  
MODEL EVALUATION OF MODELS WITH REGARD TO THE PROPOSED  
FRAMEWORK.

Model	Aug.	Easy (n=69)	Medium (n=24)	Hard (n=27)	Average (n=120)
gemma-3-27b-it	None	0.414	0.407	0.350	0.400
		[0.364, 0.467]	[0.328, 0.492]	[0.270, 0.439]	[0.362, 0.440]
	Text	0.780	0.711	0.567	0.722
		[0.733, 0.820]	[0.629, 0.781]	[0.477, 0.652]	[0.684, 0.756]
	Multi	<u>0.786</u>	<u>0.741</u>	<u>0.608</u>	<u>0.740</u>
	-modal	[0.739, 0.826]	[0.661, 0.808]	[0.519, 0.691]	[0.703, 0.774]
gpt-4o-mini	ColPali	0.548	0.459	0.458	0.510
		[0.495, 0.600]	[0.377, 0.543]	[0.372, 0.547]	[0.470, 0.550]
	None	0.568	0.422	0.325	0.487
		[0.515, 0.619]	[0.342, 0.507]	[0.248, 0.413]	[0.447, 0.527]
	Text	0.745	0.600	0.525	0.668
		[0.696, 0.788]	[0.516, 0.679]	[0.436, 0.612]	[0.630, 0.705]
gpt-4o	Multi	<u>0.754</u>	0.563	<u>0.633</u>	<u>0.687</u>
	-modal	[0.705, 0.796]	[0.479, 0.644]	[0.544, 0.714]	[0.648, 0.723]
	ColPali	0.609	<u>0.667</u>	0.500	0.600
		[0.556, 0.659]	[0.583, 0.741]	[0.412, 0.588]	[0.560, 0.638]
	None	0.664	0.630	0.358	0.595
		[0.612, 0.712]	[0.546, 0.707]	[0.278, 0.447]	[0.555, 0.634]
gemma-3-27b	Text	0.832	0.830	0.583	0.782
		[0.789, 0.868]	[0.757, 0.884]	[0.494, 0.668]	[0.747, 0.813]
	Multi	<b>0.835</b>	<b>0.874</b>	<b>0.658</b>	<b>0.808</b>
	-modal	[0.792, 0.870]	[0.807, 0.921]	[0.570, 0.737]	[0.775, 0.838]
	ColPali	0.748	0.837	0.633	0.745
		[0.699, 0.791]	[0.765, 0.891]	[0.544, 0.714]	[0.709, 0.778]

Note: [95% CI] - Agresti-Coull 95% confidence interval; underline - best performing augmentation for open-source model (gemma3-27b-it); bold best performing augmentation for proprietary model (gpt-4o); Aug. - augmentation.

In the first main experiment, gpt-4o achieved the highest overall accuracy with multi-modal augmentation (0.808; 95% CI [0.775, 0.838]) (Table II). Text augmentation (0.782; 95% CI [0.747, 0.813]) and ColPali (0.745; 95% CI [0.709, 0.778]) were slightly lower; pairwise differences were not statistically significant for gpt-4o. For gemma-3-27b-it, multi-modal (0.740; 95% CI [0.703, 0.774]) and text (0.722; 95% CI [0.684, 0.756]) significantly outperformed ColPali (0.510; 95% CI [0.470, 0.550]). By difficulty, gpt-4o with ColPali (0.745; 95% CI [0.709, 0.778]) was comparable to gemma-3-27b-it with its best augmentation (0.740; 95% CI [0.703, 0.774]).

Within-model comparisons (Table III) show that text and multi-modal augmentations outperform no augmentation and ColPali in most cases (all  $p < 0.001$ ), except for gpt-4o, where all three augmentations simply outperform no augmentation. For gemma-3-27b-it, ColPali did not outperform no augmentation.

In the second experiment (Table IV) comparing visual retrievers with the gpt-5 family, the highest overall mean accuracy was tied between ColPali (0.828; 95% CI [0.796, 0.856])

TABLE III  
PAIRWISE COMPARISONS OF STATISTICALLY SIGNIFICANT DIFFERENCE  
BETWEEN AUGMENTATION AND MODELS ( $p < 0.05$ ).

Model	Augmentations	p-value
<b>gemma-3-27b-it</b>	None < Text	< 0.001
	None < Multi-modal	< 0.001
	ColPali < Text	< 0.001
	ColPali < Multi-modal	< 0.001
<b>gpt-4o-mini</b>	None < Text	< 0.001
	None < Multi-modal	< 0.001
	None < ColPali	0.001
	ColPali < Text	0.004
<b>gpt-4o</b>	ColPali < Multi-modal	< 0.001
	None < Text	< 0.001
	None < Multi-modal	< 0.001
	None < ColPali	< 0.001

TABLE IV  
PERFORMANCE OF VISUAL RETRIEVERS (COLPALI, COLQWEN,  
COLFLOR) ACROSS THE GPT-5 MODEL FAMILY. VALUES ARE MEAN  
ACCURACY WITH AGRESTI-COULL 95% CIs.

LLM model	Retrieval model	Easy (n=69)	Medium (n=24)	Hard (n=27)	Average (n=120)
<b>gpt-5</b>	ColPali	<u>0.835</u>	0.830	0.808	<b>0.828</b>
		[0.792, 0.870]	[0.757, 0.884]	[0.728, 0.869]	<b>[0.796, 0.856]</b>
	ColQwen	<u>0.835</u>	0.822	0.783	0.822
		[0.792, 0.870]	[0.748, 0.878]	[0.701, 0.848]	[0.789, 0.850]
<b>gpt-5-mini</b>	ColFlor	0.823	0.837	<u>0.833</u>	<b>0.828</b>
		[0.779, 0.860]	[0.765, 0.891]	[0.756, 0.890]	<b>[0.796, 0.856]</b>
	ColPali	0.817	<u>0.859</u>	0.717	0.807
		[0.773, 0.855]	[0.790, 0.909]	[0.630, 0.790]	[0.773, 0.836]
<b>gpt-5-nano</b>	ColQwen	0.817	0.830	0.708	0.798
		[0.773, 0.855]	[0.757, 0.884]	[0.621, 0.782]	[0.764, 0.829]
	ColFlor	0.823	<u>0.859</u>	0.758	<u>0.818</u>
		[0.779, 0.860]	[0.790, 0.909]	[0.674, 0.827]	[0.785, 0.847]
<b>gpt-5-nano</b>	ColPali	0.797	0.763	0.600	<u>0.750</u>
		[0.751, 0.836]	[0.684, 0.827]	[0.511, 0.683]	[0.714, 0.783]
	ColQwen	0.786	0.793	0.575	0.745
		[0.739, 0.826]	[0.716, 0.853]	[0.486, 0.660]	[0.709, 0.778]
<b>gpt-5-nano</b>	ColFlor	0.759	0.770	0.592	0.728
		[0.712, 0.802]	[0.692, 0.834]	[0.502, 0.675]	[0.691, 0.762]

and ColFlor (0.828; 95% CI [0.796, 0.856]) using gpt-5. For gpt-5-mini, medium-difficulty items reached the highest category-wise accuracy with ColPali and ColFlor (both 0.859; 95% CI [0.790, 0.909]). In contrast, gpt-5-nano underperformed the larger variants by  $\sim 8$ –10 percentage points.

Within each gpt-5 model, differences among visual retrievers were small and not statistically significant. The smallest unadjusted  $p$ -value was 0.430 (ColPali vs. ColFlor on gpt-5-nano); after Bonferroni correction across within-model retriever comparisons,  $p = 1.0$ .

## V. DISCUSSION

Multi-modal RAG and OCR-free visual retrieval aim at the same goal: grounding generation on document evidence rather than model memory. Both decouple knowledge from weights and benefit from fine-grained, late-interaction matching. Where they differ is *when* visual content is inter-

preted. Modality-converting pipelines interpret earlier (via OCR/layout analysis and summarization). Visual retrievers interpret later (the LLM reads retrieved page images). This placement matters in practice.

*Experiment 1 (capacity  $\times$  pipeline).* When model capacity is limited, direct visual retrieval underperforms. In our initial screen, gemma-3-27B-IT did poorly with ColPali compared with text or multi-modal conversion. This shows a failure mode: the system can retrieve the right page but the generator cannot reliably read it. With a stronger model (GPT-4o), the gap closes and ColPali becomes competitive. The design lesson is conservative: conversion to text lowers the burden on the generator; OCR-free pipelines lean on LLM visual reasoning.

*Experiment 2 (retriever  $\times$  model family).* Within the GPT-5 family, retriever choice mattered less than model scale. ColPali, ColQwen, and ColFlor delivered near-identical means with overlapping CIs. ColFlor matched ColPali while being far smaller and faster, making it an efficient default with frontier models. Using ColFlor, gpt-5 exceeded gpt-5-mini by only about 1–2 percentage points on average, while gpt-5-mini is roughly  $5\times$  cheaper per token. For many deployments, that trade-off favors gpt-5-mini with ColFlor.

*Pipeline simplicity vs. robustness.* Visual retrievers avoid OCR and fragile PDF parsing (table reconstruction, layout heuristics), simplifying ingestion and limiting parser-induced errors. The cost is a higher reliance on the LLM’s visual reasoning. Conversion-based pipelines add engineering steps and may lose detail, but generalize better to mid-sized models and yield clearer evidence traces.

*Implication.* Choose capacity-aware designs. Under ample compute or hosted APIs, OCR-free retrieval is attractive. Under tighter budgets or on-prem constraints, conversion-based multi-modal RAG remains the safer and more interpretable default.

## A. Limitations

This study has several limitations. The benchmark is small (120 MCQs) and focused on glycobiology, so broader validation in other biomedical subfields (e.g., radiology, genomics) is needed. The item mix is skewed toward “easy” cases (57.5%), which reduces power for medium and hard analyses; we treat those strata as exploratory. We analyzed one open-source model in depth (gemma-3-27B-IT) because its initial performance was superior to other options; future work should span a wider range of open-source VLMs to map scaling and architecture effects. Finally, we reported multiple-choice accuracy; adding factual consistency and evidence-use metrics would give a fuller behavioral picture.

## B. Future Research Directions

*Retriever design.* We will systematically compare modern visual retrievers—ColPali [10], ColQwen [39], and ColFlor [13]—and newer vision/video variants to see how retrieval detail (page-, patch-, or token-level) and index compression

affect accuracy, speed, and storage on visually rich biomedical articles. Prior work shows that late-interaction, OCR-free retrievers can be both simple and highly competitive, which makes them strong candidates for evidence retrieval in practice. We will evaluate on document-centric and vision-centric benchmarks to ensure results transfer beyond a single dataset.

*Lightweight and cost-aware MM-RAG.* We will pair smaller, faster retrievers with mid-sized generators to reduce cost and latency while tracking accuracy and “cost per correct answer.” ColFlor is an example of a compact, OCR-free retriever (about 174M parameters) that approaches the performance of heavier models like ColPali, but runs substantially faster, suggesting good cost–utility for institutional deployments.

*Trustworthy grounding.* Beyond accuracy, we will stress-test how reliably models use retrieved evidence—especially images, tables, and complex page layouts. Recent vision-centric studies find that even strong models under-use retrieved visuals, highlighting the need for strict citation, showing the exact snippets or page crops used, and simple checks that link each answer back to its sources. These measures support auditability in clinical and research settings.

*Grounded platform for researchers.* We intend to turn these ideas into a simple, secure platform that helps researchers in their own domain ask better questions, find the right evidence, and draft stronger proposals. The system will use RAG so every answer is *grounded* in trusted sources the user selects/provides (papers, guidelines, lab documents, PDFs—including figures and tables). For visuals, it will either convert images/tables to text or pass the page image—whichever is clearer—and always show the exact snippets used so claims can be checked. The platform will scale from a single lab to an institution, support on-premise options to protect sensitive data, and let teams choose smaller or larger models to match budget and accuracy needs.

## VI. CONCLUSION

This study compared text-centric and OCR-free visual retrieval strategies for multi-modal grounding on a glycobiology MCQ benchmark. We found a clear capacity-dependent pattern: conversion-based pipelines (text and multi-modal summaries) are more reliable with mid-size VLMs (e.g., Gemma-3-27B-IT), whereas OCR-free late-interaction retrieval becomes competitive with frontier models (e.g., GPT-4o / GPT-5). Among visual-retrievers, ColFlor matched ColPali under GPT-5 while offering a smaller footprint, suggesting an attractive efficiency-accuracy trade-off. Practically, these results argue for capacity-aware MM-RAG design: convert earlier when model visual reasoning is constrained; retrieve page images directly when model capacity allows.

Our work is preliminary. The benchmark is small and single-domain. Broader evaluations across modalities and domain will test the generality of these findings and support trustworthy multi-modal assistants in biomedicine.

## GENAI USAGE DISCLOSURE

In this manuscript we used Gemini and ChatGPT to help improve its language and overall presentation. Our aim with these tools was to enhance the text’s clarity, coherence, and correctness, and we carefully reviewed and edited any suggestions they provided. We want to clearly state that all scientific content, core ideas, and analyses are entirely our own original work, developed without relying on Gemini or ChatGPT for any conceptual or analytical aspects of the research. As authors, we are responsible for the manuscript’s final content and affirm the originality of its scientific contributions stemming from our own efforts.

## CODE AVAILABILITY

A reproducibility package (benchmark template, scripts and prompts) are available at: “[https://github.com/pkocbek/multi-modal\\_colpali](https://github.com/pkocbek/multi-modal_colpali)”.

## REFERENCES

- [1] K. Singhal, S. Azizi, T. Tu, and et al., “Large language models encode clinical knowledge,” *Nature*, vol. 620, no. 7974, pp. 172–180, 2023.
- [2] G. Tsatsaronis, G. Balikas, P. Malakasiotis, and et al., “An overview of the bioasq large-scale biomedical semantic indexing and question answering competition,” *BMC Bioinformatics*, vol. 16, p. 138, 2015.
- [3] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, “Pubmedqa: A dataset for biomedical research question answering,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 2567–2577. [Online]. Available: <https://aclanthology.org/D19-1259/>
- [4] O. Ovadia, M. Brief, M. Mishaeli, and O. Elisha, “Fine-tuning or retrieval? comparing knowledge injection in llms,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.15/>
- [5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. tau Yih, T. Rockt’aschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.
- [6] M. Editors, “Retrieval-augmented generation in biomedical informatics,” *Biomedicine*, vol. 12, no. 7, p. 687, 2024. [Online]. Available: <https://www.mdpi.com/2306-5354/12/7/687>
- [7] M. Frank and S. Schloissnig, “Bioinformatics and molecular modeling in glycobiology,” *Cellular and Molecular Life Sciences*, vol. 67, no. 16, pp. 2749–2772, 2010.
- [8] D. O. Williams and E. Fadda, “Can chatgpt pass glycobiology?” *Glycobiology*, vol. 33, no. 8, pp. 606–614, 2023. [Online]. Available: <https://academic.oup.com/glycob/article/33/8/606/7235670>
- [9] M. Alkhalaf, P. Yu, M. Yin, and C. Deng, “Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records,” *Journal of Biomedical Informatics*, vol. 156, p. 104662, 2024.
- [10] M. Faysse, A. Loison, Q. Macé et al., “Colpali: Efficient document retrieval with vision language models,” in *International Conference on Learning Representations (ICLR)*, 2025. [Online]. Available: <https://arxiv.org/abs/2407.01449>
- [11] O. Khattab and M. Zaharia, “Colbert: Efficient and effective passage search via contextualized late interaction over bert,” in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 39–48.
- [12] S. Yu, C. Tang, B. Xu, J. Cui, J. Ran, Y. Yan, Z. Liu, S. Wang, X. Han, Z. Liu, and M. Sun, “Visrag: Vision-based retrieval-augmented generation on multi-modality documents,” in *International Conference on Learning Representations (ICLR)*, 2025, iCLR 2025. [Online]. Available: <https://openreview.net/forum?id=zG459X3Xge>

- [13] A. Masry and et al., “Colflor: Towards bert-size vision-language document retrieval models,” *arXiv preprint arXiv:2405.05666*, 2024. [Online]. Available: <https://huggingface.co/blog/ahmed-masry/colflor>
- [14] V. Team, “Colqwen2-v0.2: Vision-language document retrieval,” <https://huggingface.co/vidore/colqwen2-v0.2>, 2025.
- [15] Google Research, “Multimodal medical ai,” 2023, accessed September 29, 2025. [Online]. Available: <https://research.google/>
- [16] Y. Li and et al., “A comprehensive review of ai in multimodal biomedical data analysis,” *Genomics, Proteomics & Bioinformatics*, vol. 23, no. 1, p. qzaf011, 2024.
- [17] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, “Llava-med: Training a large language-and-vision assistant for biomedicine in one day,” in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023. [Online]. Available: <https://aka.ms/llava-med>
- [18] M. Moor, S. Pineda, K. Kreis, M. Welling, M. Kraus, D. Rueckert, and G. R’atsch, “Med-flamingo: a multimodal medical few-shot learner,” in *Proceedings of the Machine Learning for Health (ML4H) 2023*, vol. 225. PMLR, 2023, pp. 353–367. [Online]. Available: <https://proceedings.mlr.press/v225/moor23a.html>
- [19] T. Tu and et al., “Towards generalist biomedical ai,” *arXiv preprint arXiv:2307.14334*, 2023. [Online]. Available: <https://arxiv.org/abs/2307.14334>
- [20] S. Liu, A. B. McCoy, and A. Wright, “Improving large language model applications in biomedicine with retrieval-augmented generation: A systematic review, meta-analysis, and clinical development guidelines,” *Journal of the American Medical Informatics Association*, vol. 32, no. 4, pp. 605–615, 2025.
- [21] M. F. Aljunid and et al., “Retrieval-augmented generation (rag) in healthcare: A comprehensive review,” *AI*, vol. 6, no. 9, p. 226, 2025.
- [22] Emergent Mind Team, “Biomedical literature q&a system using retrieval-augmented generation (rag),” *arXiv preprint arXiv:2509.05505*, 2025. [Online]. Available: <https://arxiv.org/abs/2509.05505>
- [23] Emergent Mind, “Biomedical literature q&a system using retrieval-augmented generation (rag),” 2025. [Online]. Available: <https://www.emergentmind.com/biomedragqa>
- [24] V. S. Sree and et al., “Medgpt: A modular chatbot for medical diagnostics,” in *Proceedings of the 2024 International Conference on Intelligent Systems and Machine Learning (ISML)*, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/medgpt2024>
- [25] W. Hu and et al., “Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models,” in *International Conference on Learning Representations (ICLR)*, 2025. [Online]. Available: <https://iclr.cc/virtual/2025/poster/2410.08182>
- [26] D. Cheng, S. Huang, Z. Zhu, X. Zhang, W. X. Zhao, Z. Luan, B. Dai, and Z. Zhang, “On domain-specific post-training for multimodal large language models,” *arXiv preprint arXiv:2411.19930*, 2024. [Online]. Available: <https://arxiv.org/abs/2411.19930>
- [27] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, and et al., “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution,” *arXiv preprint arXiv:2409.12191*, 2024. [Online]. Available: <https://arxiv.org/abs/2409.12191>
- [28] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, “Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models,” *arXiv preprint arXiv:2407.07895*, 2024. [Online]. Available: <https://arxiv.org/abs/2407.07895>
- [29] Meta AI, “Llama 3.2: Revolutionizing edge ai and vision with open, customizable models,” 2024. [Online]. Available: <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>
- [30] A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, and et al., “Gemma 3 technical report,” *arXiv preprint arXiv:2503.19786*, 2025. [Online]. Available: <https://arxiv.org/abs/2503.19786>
- [31] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, “Efficient memory management for large language model serving with pagedattention,” in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- [32] OpenAI, “Enterprise privacy at openai,” 2025. [Online]. Available: <https://openai.com/enterprise-privacy/>
- [33] C. Auer, M. Lysak, A. Nassar, M. Dolfi, N. Livathinos, P. Vagenas, C. Berrospi Ramis, M. Omenetti, F. Lindlbauer, K. Dinkla, and et al., “Docling technical report,” *arXiv preprint arXiv:2408.09869*, 2024. [Online]. Available: <https://arxiv.org/abs/2408.09869>
- [34] Qdrant Team, “Qdrant: Open-source vector database and vector search engine,” 2025. [Online]. Available: <https://qdrant.tech/>
- [35] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie, “C-pack: Packed resources for general chinese embeddings,” in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 1–9. [Online]. Available: <https://doi.org/10.1145/3626772.3657878>
- [36] Q. Macé, A. Loison, and M. Faysse, “Vidore benchmark v2: Raising the bar for visual retrieval,” *arXiv preprint arXiv:2505.17166*, 2025. [Online]. Available: <https://arxiv.org/abs/2505.17166>
- [37] “vidore/colpali-v1.3-merged,” <https://huggingface.co/vidore/colpali-v1.3-merged>, 2025, model card of ColPali (2.92B parameters).
- [38] “Colpali model documentation in hugging face transformers,” [https://huggingface.co/docs/transformers/en/model\\_doc/colpali](https://huggingface.co/docs/transformers/en/model_doc/colpali), 2025.
- [39] “manu/colqwen2-v0.2,” <https://huggingface.co/manu/colqwen2-v0.2>, 2025, model card of ColQwen2 (adapter over Qwen2-VL).
- [40] “ahmed-masry/colflor model card,” <https://huggingface.co/ahmed-masry/ColFlor>, 2024, 174M-parameter OCR-free visual retriever.
- [41] “Colflor: Towards bert-size vision-language document retrieval models,” <https://huggingface.co/blog/ahmed-masry/colflor>, 2024, details on architecture, speeds, and performance.
- [42] A. Agresti and B. A. Coull, “Approximate is better than “exact” for interval estimation of binomial proportions,” *The American Statistician*, vol. 52, no. 2, pp. 119–126, 1998.

## APPENDIX A

### LIST OF DOI LINKS FOR THE 25 ORIGINAL RESEARCH AND REVIEW MANUSCRIPTS

<https://doi.org/10.1186/s12967-018-1695-0>  
<https://doi.org/10.1097/hjh.0000000000002963>  
<https://doi.org/10.1186/s12967-018-1616-2>  
<https://doi.org/10.3390%2Fbiom13020375>  
<https://doi.org/10.1016/j.bbagen.2017.06.020>  
<https://doi.org/10.1172%2Fjci.insight.89703>  
<https://doi.org/10.1016%2Fj.isci.2022.103897>  
<https://doi.org/10.1002%2Fart.39273>  
<https://doi.org/10.1016/j.bbadis.2018.03.018>  
<https://doi.org/10.1097/MIB.0000000000000372>  
<https://doi.org/10.1053%2Fj.gastro.2018.01.002>  
<https://doi.org/10.1186/s13075-017-1389-7>  
<https://doi.org/10.1021/pr400589m>  
<https://doi.org/10.1161/CIRCRESAHA.117.312174>  
<https://doi.org/10.2337/dc22-0833>  
<https://doi.org/10.1097%2FMD.0000000000003379>  
<https://doi.org/10.1158/1078-0432.CCR-15-1867>  
<https://doi.org/10.1093/gerona/glt190>  
<https://doi.org/10.1111/imr.13407>  
<https://doi.org/10.1053/j.gastro.2018.05.030>  
<https://doi.org/10.1016/j.csbj.2024.03.008>  
<https://doi.org/10.1016/j.cellimm.2018.07.009>  
<https://doi.org/10.1016/j.biotechadv.2023.108169>  
<https://doi.org/10.4049/jimmunol.2400447>

## APPENDIX B

### GENERAL PROMPT TEMPLATE WAS USED FOR CREATING TABLE AND FIGURE SUMMARIES

You are an AI assistant specialized in summarizing tables and figures for efficient retrieval. \n\nInstructions: \n\nIdentify Input Type: Explicitly state whether the input provided is a table or a figure.\n\nScientific Abstract: Summarize the contents concisely in the style of a scientific abstract. Include relevant numeric values and key findings. \n\nRetrieval Optimization: Structure your summary clearly, optimizing keywords and phrasing to enhance retrieval and indexing.

\nLength Constraint: Your summary must strictly adhere to a maximum of 300 words or 250 tokens. Do not exceed this limit under any circumstances. Any text exceeding will be just cutoff post generation.\nAvoid Generic Openings: Do not start your summary with generic phrases such as "The image provided is," "The table shows," or similar introductory sentences. Instead, immediately describe the core content.\nPrevent Redundancy: Write succinctly, avoiding repetition of concepts or data points.\nFinal output: Only summary text. If no relevant data is present, output \'\'.

Additional information if provided in the Context below. If the Context is not empty, analyse it and choose from the letters. MAKE SURE your output is one of the four values stated. Here is the query: {question}. Here are the choices: {question\_string} Context:

## APPENDIX C

### EXAMPLE DOCKER CONFIGURATIONS FOR DOCLING, QDRANT AND GEMMA-3-27B-IT

#### Dockling:

```
docker run \
  --rm --gpus all \
  -e DOCLING_SERVE_ENABLE_UI=true \
  -e DOCLING_SERVE_MAX_SYNC_WAIT=600 \
  -e DOCLING_SERVE_ENABLE_REMOTE_SERVICES=true \
  -e "DOCLING_SERVE_API_KEY=${DOCLING_API_KEY}" \
  --name docling_serve \
  -p 5001:5001 \
  ghcr.io/docling-project/docling-serve-cul24:latest
```

#### Qdrant:

```
docker run \
  --rm \
  --name qdrant_vd \
  --gpus=all \
  -p 6333:6333 \
  -p 6334:6334 \
  -e QDRANT_GPU_INDEXING=1 \
  -e "QDRANT_SERVICE_API_KEY=${QDRANT_API_KEY}" \
  --ulimit nofile=65536:65536 \
  -v ./src/vectordb/storage:/qdrant/storage \
  -v ./src/vectordb/custom_config.yaml:
  /qdrant/config/custom_config.yaml \
  qdrant/qdrant:gpu-nvidia-latest
```

#### Gemma-3-27b-it (vLLM):

```
docker run --gpus all -it --rm --pull=always \
  --name gemma_27b \
  -v "${HF_DIR}:/root/.cache/huggingface" \
  --env "HUGGING_FACE_HUB_TOKEN=
  ${HUGGING_FACE_HUB_TOKEN}" \
  --env TRANSFORMERS_OFFLINE=1 \
  --env VLLM_RPC_TIMEOUT=180000 \
  --env HF_DATASET_OFFLINE=1 \
  -p 8006:8000 \
  --ipc=host \
  vllm/vllm-openai:latest \
  --model "google/gemma-3-27b-it" \
  --limit_mm_per_prompt '{"image": 8}' \
  --gpu-memory-utilization 0.82 \
  --max_model_len 16000 \
  --enable-sleep-mode
```

## APPENDIX D

### GENERAL PROMPT TEMPLATE FOR EVALUATION

Generate a JSON with the query\_answer, the answer provided behind the letters: A, B, C, and D. These are the values.