

Graph Retrieval in Multimodal RAG for Biomedical Documents: A Pharmacokinetics and Pharmacodynamics Case Study

Primoz Kocbek^{1,2} Katarina Mestrovic¹ Nguyen Thu Huong³
Ales Fajmut⁴ Gregor Stiglic^{1,5}

¹University of Maribor, Faculty of Health Sciences ²University of Ljubljana, Faculty of Medicine

³Phan Chau Trinh University ⁴University of Maribor, Faculty of Natural Sciences and Mathematics

⁵Usher Institute, The University of Edinburgh

{primoz.kocbek, katarina.mestrovic2, ales.fajmut, gregor.stiglic}@um.si huong.nt@pctu.edu.vn

Abstract

Introduction: Biomedical PDFs distribute answer evidence across prose, tables, figures, and cross-document relations, complicating the retrieval of answer-sufficient context. We examine whether graph and multimodal retrieval provide complementary answers and whether direct concatenation exploits this complementarity. **Methods:** We screen 13 local generators on 205 pharmacokinetics/pharmacodynamics multiple-choice questions and evaluate 11 retrieval conditions over ten runs for five selected models. We quantify questions answered correctly by either or both retrieval routes and compare direct fusion with each standalone route using paired question-level bootstrap intervals. A separate ten-run panel tests answer order, irrelevant context, paraphrasing, generated distractors, and option removal. **Results:** Multimodal retrieval alone answers 11.5–16.2% of questions correctly that graph retrieval misses, while graph retrieval alone answers 7.9–13.9% that multimodal retrieval misses. Counting a question as correct whenever either route succeeds yields an oracle upper bound of 69.5–77.7%, exceeding direct concatenation by 8.4–13.5 percentage points. Direct fusion improves on the stronger standalone route by at most 0.6 points; paired intervals include zero in all ten model–depth comparisons. In repeated evaluation, Gemma-4-31B with multimodal retrieval achieved the highest accuracy (66.6%, 95% CI 65.4–67.8). **Discussion:** The central finding is substantial answer-level complementarity that direct concatenation does not reliably exploit. The oracle bound measures retrospective selection, not an implemented routing method. These findings motivate evaluating source-aware routing or reranking to select between complementary retrieval routes.

1 Introduction

Retrieval-augmented generation (RAG) grounds model outputs in external evidence (Lewis et al.,

2020), but biomedical PDFs complicate evidence retrieval. A single answer may depend on prose that identifies a drug or endpoint and a table or figure that provides the decisive dose, population, or numerical value. Document-level recall is therefore insufficient: the system must retrieve answer-sufficient evidence while preserving its source and page context.

Our previous study in glycobiology compared text conversion with OCR-free page retrieval on 120 questions (Kocbek et al., 2025). Its results showed that retrieval effectiveness depended on both evidence representation and generator capability. The present PK/PD study shifts the focus from document conversion to evidence selection: once text, tables, figures, and pages are indexed, does graph retrieval improve biomedical question answering, and can graph and multimodal evidence be combined effectively?

Graph retrieval can connect entities and relations that are separated across chunks or documents. LightRAG integrates graph and vector representations with low- and high-level retrieval (Guo et al., 2025b); our pipeline extends its input with source-linked text, tables, figures, and image descriptions, consistent with the multimodal design of RAG-Anything (Guo et al., 2025a). This expansion may recover otherwise fragmented evidence, but it may also introduce generic entities, indirect relations, or excessive context that displaces the exact passage, table cell, or figure needed for an answer.

Figure 1 summarizes the evidence-first architecture, which uses MinerU for PDF extraction (Wang et al., 2024), Milvus for source-linked search (Wang et al., 2021), and model-isolated LightRAG graphs. Across thirteen local generators, we compare text, visual-asset, multimodal, page-level late-interaction, graph, and graph-multimodal retrieval. Five selected models are evaluated over ten repeated runs; additional pertur-

bations test answer order, irrelevant context, paraphrasing, distractors, and option removal. A central contribution is the answer-level complementarity analysis: graph and multimodal routes answer partly different questions correctly. Selecting the correct answer whenever either route succeeds yields an oracle upper bound 8.4–13.5 percentage points above direct concatenation.

2 Related Work

Building RAG systems for biomedical documents requires a common representation that can support several retrieval routes. Scientific PDFs must first be converted into structured, source-linked evidence that preserves prose, tables, figures, captions, and page provenance. We use MinerU because a single extraction pipeline can be reused for text retrieval, multimodal ingestion, and LightRAG graph construction: text and Markdown support chunking and graph extraction, while tables, figures, crops, and captions provide typed multimodal evidence (Wang et al., 2024). This shared preprocessing avoids separate document conversions for each retriever and keeps their outputs aligned to the same source locations.

The resulting representations also require a storage layer that can serve different indexing strategies. Milvus provides scalable vector management and similarity search (Wang et al., 2021) while separating storage, indexing, and query execution (Guo et al., 2022). In our architecture, it stores dense text and visual embeddings, LightRAG chunks, entities, and relations, and the page-level multi-vector representations required by late-interaction retrievers such as ColPali and ColQwen. Document, page, modality, retrieval-model, and generator identifiers preserve provenance and isolate model-specific collections while allowing all retrieval routes to use a common backend.

Multimodal retrieval can index extracted assets or rendered pages. ColPali extends ColBERT-style late interaction (Khattab and Zaharia, 2020) by matching query tokens against page-patch embeddings (Faysse et al., 2025). This preserves localized visual signals without cross-encoding every page, but increases storage and similarity cost; reducing page vectors trades efficiency against quality (Ma et al., 2025). Joint image–text memory can outperform text-only retrieval (Chen et al., 2022), yet realistic multimodal systems remain vulnera-

ble to tables and rephrased queries (Wasserman et al., 2025). A highly ranked page is therefore not necessarily answer-sufficient evidence.

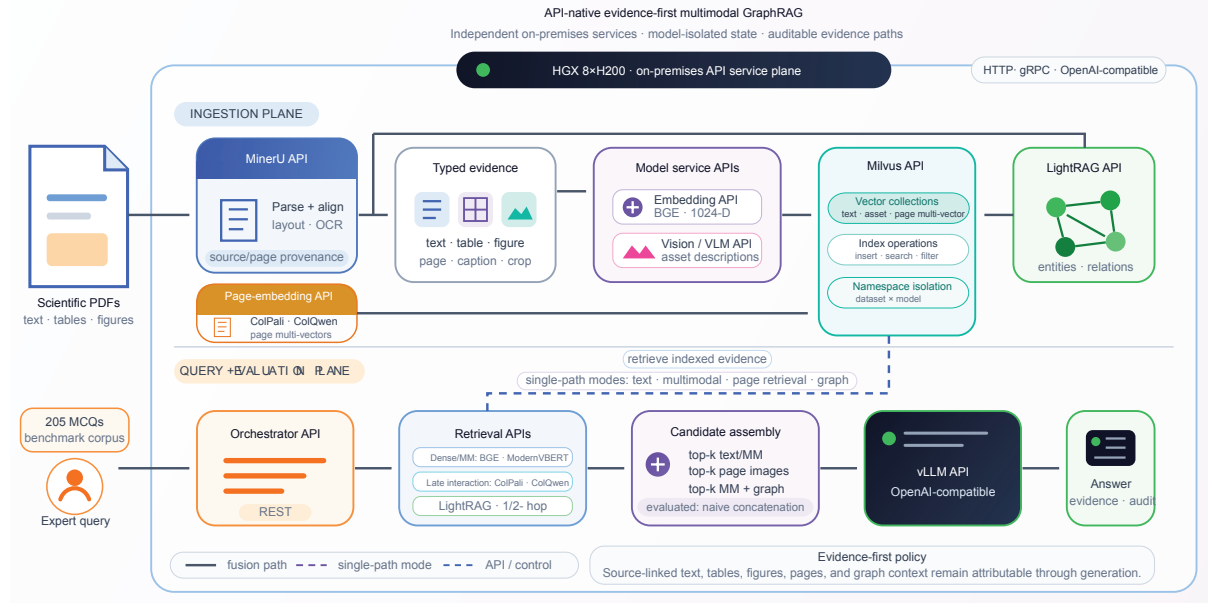
Graph retrieval represents relations spanning chunks, pages, or documents. GraphRAG uses entity graphs and community summaries (Edge et al., 2024); LightRAG combines graph indexing, vectors, dual-level retrieval, and incremental updates (Guo et al., 2025b). Its RAG-Anything integration supports interleaved text, tables, images, and equations (Guo et al., 2025a). We ask whether graph structure adds decisive biomedical evidence and whether it can be combined with multimodal candidates without indirect context displacing direct evidence.

Multiple-choice questions (MCQs) permit consistent scoring across models and runs without a generative judge, but options can reveal cues, support elimination, and make results sensitive to order and distractor quality (Balepur et al., 2025). Surface validity is distinct from psychometric plausibility (Feng et al., 2024), and choice-aware generation can produce more discriminating alternatives than generic methods (Lee et al., 2025). We therefore treat MCQ accuracy as an evaluation instrument, and audit answer order, paraphrasing, distractor replacement, and option removal.

3 Materials and Methods

3.1 Benchmark and task

The benchmark contains 205 PK/PD MCQs: 68 easy, 69 medium, and 68 hard. A master’s-trained domain researcher developed and refined the questions with input from a doctoral student, using 60 source papers. The benchmark workbook records source DOIs and answer locations in pages, tables, figures, or text. Multiple development iterations reviewed the questions and retrieval-difficulty labels: when a model correctly answered a hard question with sufficient supporting explanation, the question was downgraded or replaced. This constitutes iterative domain review; independent expert validation remains a future step. It covers antimicrobial PK/PD, with particular emphasis on beta-lactam therapy. Questions address dosing and infusion strategies, exposure targets, therapeutic drug monitoring, tissue distribution, efficacy and toxicity, and dose individualization in clinically complex populations. Easy questions use explicit textual evidence; medium questions require tables, figures, or within-document synthesis; hard ques-



Note: MinerU produces source-linked text, tables, figures, captions, crops, and page images. Independent services create dense, multimodal, and page-level multi-vector representations. Milvus manages model-isolated collections, while LightRAG maintains entity-relation retrieval. At query time, the orchestrator invokes individual retrieval routes or their evaluated concatenations before evidence-grounded generation through an OpenAI-compatible vLLM endpoint.

Figure 1: API-native, evidence-first multimodal GraphRAG architecture and evaluation workflow.

tions integrate sections or papers. Options are included in the retrieval query because a distinguishing drug, dose, endpoint, or value may appear only there. The primary outcome is exact option accuracy. Document hit rate indicates whether the linked paper was retrieved, not whether the context was answer sufficient.

3.2 Parsing and evidence preservation

MinerU converts PDFs into structured Markdown and visual assets. Documents longer than 20 pages are rendered and processed in consecutive 20-page batches to control memory use and processing time. The MinerU outputs are then reassembled in their original page order before ingestion, with global page numbers and source provenance preserved across batch boundaries. Typed table records retain body, caption, footnote, crop, source, and page; figure records retain crop, caption, local text, source, and page. This follows MinerU’s structured extraction (Wang et al., 2024) and RAG-Anything’s interleaved multimodal representation (Guo et al., 2025a).

Text chunks and visual descriptions are embedded and stored in Milvus (Guo et al., 2022). Approximate search uses a Hierarchical Navigable Small World (HNSW) index, a multilayer proximity graph that efficiently narrows the search from broad neighborhoods to nearby vectors (Malkov and Yashunin, 2020); page-image retrievers use

ColPali-style late interaction (Faysse et al., 2025).

3.3 Native LightRAG graph

LightRAG constructs graphs from MinerU Markdown and typed multimodal records (Guo et al., 2025b). LLM-based indexing extracts domain entities and relations; query keywords select entities and higher-level concepts. Each dataset-model pair uses a separate Milvus database to isolate graphs, vectors, and assets.

Context prioritizes concrete text windows, table bodies, figure-local text, source links, and page metadata over graph summaries. We evaluate one- and two-hop traversal. Combined conditions concatenate LightRAG nodes and relations with the top 10 multimodal candidates, without learned gating or reranking (Section 5.1). Graph-MM 1-Hop and Graph-MM 2-Hop denote these direct-concatenation baselines.

3.4 Retrieval conditions and models

We evaluate zero-shot, text, visual-asset, and multimodal RAG; ColPali, ColQwen, and ModernVBERT page retrieval; graph retrieval at one and two hops; and graph-multimodal retrieval at both depths. The 13-model screen spans dense and sparse mixture-of-experts generators with 4B–428B total parameters. Models are served locally with vLLM 0.24.0 on one to four H200 GPUs, using checkpoint-native BF16, FP8, NVFP4, INT4,

W4A4, or MXFP8 formats and context windows of 110,000–262,144 tokens (Kwon et al., 2023). Sampling follows checkpoint recommendations; the five repeated models use temperature 1.0 and top- $p = 0.95$, with top- $k = 64$ for Gemma, 20 for Qwen, and no explicit top- k for MiniMax-M3. Retrieval uses top- $k = 10$. The random seed was unset; the runtime default is presumed to be 0.

3.5 Screening and repeated analysis

In Stage 1, we screen 13 generators once across zero-shot, text, multimodal, graph, and graph-multimodal conditions. Accuracy, model size, and local feasibility are then used to select five models: Gemma-4-31B, Gemma-4-26B-A4B, Qwen3.6-27B, Qwen3.6-35B-A3B, and MiniMax-M3. In Stage 2, we repeat all 11 conditions ten times for these models. Stages 1 and 2 use exactly the same 205 questions, with no development/test split. Each completed run is one statistical sample ($n = 10$ runs per model-condition), and 95% intervals are calculated across the ten run-level metric estimates to quantify run-to-run variability. For Gemma-4-31B, we also aggregate accuracy, total tokens, and hit rate within each run and difficulty stratum.

Within each run, we classify each question as correct under both routes, multimodal only, graph only, or neither. The union of the first three defines an ideal-selection upper bound; route-only categories measure complementarity, and its gap from direct fusion measures unrealized potential. Complementarity estimates are then summarized across the same ten run-level samples ($n = 10$). We additionally calculate paired 95% question-cluster bootstrap intervals for direct fusion minus each standalone route, by resampling questions 10,000 times. Each sampled question retains its results from all ten runs and all compared methods; duplicate questions are sampled together. This upper bound uses oracle selection: it chooses the successful route after the correct answer is known. It therefore measures the potential value of route selection, but not the performance of a deployable system that must choose a route before answering.

3.6 Robustness perturbations

In Stage 3, we evaluate the five selected generators under zero-shot and text retrieval over ten runs per condition. Answer shuffling probes positional bias (Pezeshkpour and Hruschka, 2024); context noise probes distractibility (Shi et al., 2023; Liu

et al., 2024); and paraphrasing probes semantic invariance and can support decontamination audits (Yang et al., 2023). Generated distractors replace incorrect options, while the open-ended condition removes all options (Feng et al., 2024; Lee et al., 2025; Balepur et al., 2025).

We pool binary responses to obtain 95% Agresti–Coull intervals, which estimate a binomial proportion using adjusted success counts and an adjusted effective sample size, providing more stable coverage than the simple Wald interval (Agresti and Coull, 1998). Each perturbation is compared with its baseline using a two-proportion z-test; baselines have no test. Results are exploratory because pooling ignores question pairing and repeated-run dependence.

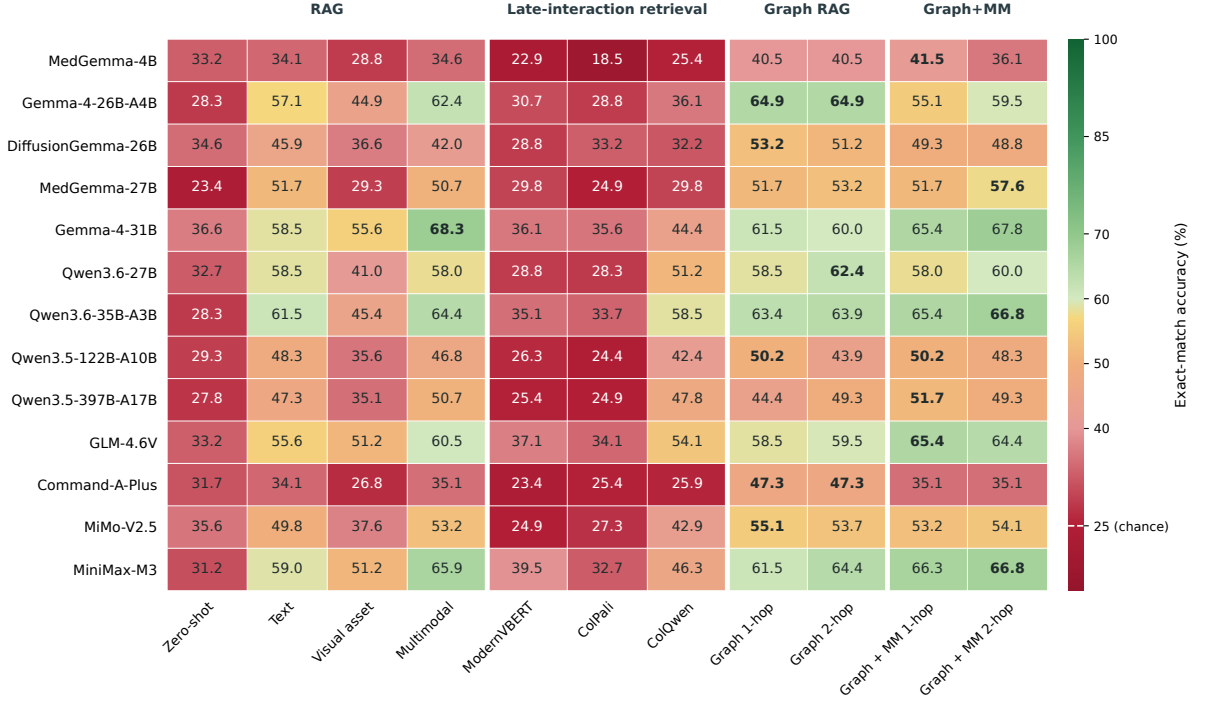
4 Results

4.1 Stage 1: broad single-run screen

Figure 2 shows that no retrieval architecture dominated the broad 13-generator screen. Gemma-4-31B was strongest with multimodal retrieval (68.3%), Gemma-4-26B-A4B with both one- and two-hop graph retrieval (64.9% for each), and Qwen3.6-35B-A3B with two-hop graph-multimodal fusion (66.8%). ColQwen raised Qwen3.6-27B page retrieval to 51.2%, compared with 28.3% for ColPali, but remained below that model’s best graph result of 62.4%. This heterogeneity motivated selection of five models for repeated evaluation in Stage 2.

4.2 Stage 2: retrieval analysis (n=10)

Table 2 compares the 11 repeated conditions. Gemma-4-31B is highest with multimodal retrieval (66.6%, 95% CI 65.4–67.8), followed by Qwen3.6-27B with multimodal retrieval (65.1%, 63.1–67.2) and MiniMax-M3 with one-hop direct fusion (65.0%, 62.8–67.3). Direct fusion exceeds the better same-depth component only for Qwen3.6-35B-A3B (+0.6 and +0.1 percentage points at one and two hops) and MiniMax-M3 at one hop (+0.5 percentage points). Paired question-cluster intervals for direct fusion minus the stronger standalone route include zero in all ten model-depth comparisons. These exploratory results concern direct concatenation, not evidence fusion generally.



Note: Cells report exact-match accuracy on 205 four-option questions. Graph+MM denotes direct concatenation of graph and multimodal candidates; white separators group conventional, page-level, graph, and fused retrieval routes.

Figure 2: Accuracy across generators and retrieval architectures in Stage 1.

Table 2: Retrieval comparison in Stage 2 across $n = 10$ runs.

Model	Metric	Zero-shot generation	Text retrieval	Visual assets	Multi-modal	Modern-VBERT	Col-Pali	Col-Qwen	Graph 1-hop	Graph 2-hop	Graph+MM 1-hop	Graph+MM 2-hop
Gemma-4-31B	Acc.	36.2 [35.0-37.5]	60.3 [59.5-61.1]	54.5 [53.1-55.9]	66.6 [65.4-67.8]	33.7 [32.7-34.8]	36.6 [35.1-38.1]	46.6 [45.1-48.1]	61.3 [60.2-62.4]	60.6 [59.4-61.9]	65.8 [64.5-67.0]	66.1 [65.3-66.9]
	Hit	0.0 [n/a]	93.2 [93.2-93.2]	88.8 [88.8-88.8]	94.6 [94.6-94.6]	33.7 [33.7-33.7]	31.7 [31.7-31.7]	89.3 [89.3-89.3]	96.9 [96.7-97.1]	96.3 [95.9-96.7]	93.0 [92.4-93.6]	92.7 [92.3-93.1]
	Tok.	564/125 [564-564/124-125]	5,363/129 [5,363-5,363/128-130]	6,292/165 [6,292-6,292/162-167]	5,568/140 [5,568-5,568/139-141]	3,783/140 [3,783-3,783/138-142]	3,775/150 [3,775-3,775/147-153]	3,774/160 [3,774-3,774/156-163]	5,796/131 [5,792-5,809/130-133]	5,886/131 [5,876-5,895/131-132]	6,076/130 [6,063-6,090/129-131]	6,132/130 [6,118-6,146/129-131]
Gemma-4-26B-A4B	Acc.	30.7 [29.1-32.2]	57.0 [55.0-59.0]	43.8 [42.4-45.1]	60.8 [59.6-62.0]	32.7 [30.5-34.9]	33.0 [31.8-34.3]	40.1 [37.6-42.6]	59.6 [58.2-60.9]	59.9 [58.6-60.9]	58.3 [56.9-59.8]	59.1 [57.9-60.2]
	Hit	0.0 [n/a]	92.7 [92.7-92.7]	90.1 [89.9-90.3]	94.1 [94.1-94.1]	33.7 [33.7-33.7]	33.7 [33.7-33.7]	89.3 [89.3-89.3]	95.4 [95.0-95.7]	96.3 [96.0-96.3]	92.5 [92.3-92.8]	94.0 [93.7-94.2]
	Tok.	564/125 [564-564/124-125]	5,367/150 [5,367-5,367/147-153]	6,552/208 [6,541-6,562/200-217]	5,652/166 [5,652-5,652/160-170]	3,783/155 [3,783-3,783/147-163]	3,775/168 [3,775-3,775/160-175]	3,774/195 [3,774-3,774/184-200]	5,564/161 [5,555-5,572/158-165]	5,564/161 [5,555-5,569/158-165]	6,164/158 [6,149-6,178/155-165]	6,218/162 [6,202-6,235/158-166]
Qwen3.6-27B	Acc.	34.0 [31.8-36.2]	61.3 [59.7-62.8]	52.5 [50.0-55.1]	65.1 [63.1-67.2]	33.5 [31.6-35.4]	33.2 [31.4-34.9]	47.1 [45.8-48.4]	61.7 [60.2-63.2]	61.5 [60.3-62.8]	64.9 [63.3-66.5]	64.2 [63.3-66.2]
	Hit	0.0 [n/a]	91.7 [91.7-91.7]	89.8 [89.8-89.8]	94.6 [94.6-94.6]	33.7 [33.7-33.7]	33.7 [33.7-33.7]	89.3 [89.3-89.3]	97.2 [96.7-97.7]	97.2 [96.8-97.7]	93.6 [93.0-94.1]	94.0 [93.3-94.6]
	Tok.	554/190 [554-554/184-196]	5,554/285 [5,554-5,554/268-300]	9,949/299 [9,949-9,949/277-321]	6,661/269 [6,661-6,661/238-294]	8,833/224 [8,833-8,833/210-238]	8,895/247 [8,895-8,895/237-262]	8,894/298 [8,894-8,894/276-320]	6,177/263 [6,164-6,190/258-270]	6,258/265 [6,252-6,265/248-281]	6,539/263 [6,522-6,556/248-282]	6,620/298 [6,607-6,634/285-311]
Qwen3.6-35B-A3B	Acc.	33.5 [32.3-34.7]	57.6 [55.6-59.5]	42.3 [40.7-43.8]	55.8 [54.4-57.2]	30.4 [29.4-31.4]	31.1 [29.7-32.4]	37.6 [35.8-39.4]	57.6 [56.0-59.8]	58.0 [56.6-60.1]	58.2 [56.0-60.4]	58.1 [56.2-60.0]
	Hit	0.0 [n/a]	92.2 [92.2-92.2]	87.8 [87.8-87.8]	93.7 [93.7-93.7]	33.7 [33.7-33.7]	24.4 [24.4-24.4]	89.3 [89.3-89.3]	95.9 [95.4-96.5]	96.3 [95.9-96.7]	92.5 [91.7-93.3]	93.7 [93.2-94.1]
	Tok.	554/141 [554-554/141-141]	5,560/258 [5,560-5,560/234-281]	10,798/192 [10,798-10,798/183-201]	8,866/170 [8,866-8,866/159-181]	8,833/149 [8,833-8,833/143-154]	8,895/169 [8,895-8,895/157-182]	8,894/170 [8,894-8,894/158-183]	6,251/213 [6,241-6,262/195-232]	6,306/228 [6,292-6,321/205-251]	6,728/188 [6,710-6,746/166-210]	6,776/177 [6,764-6,788/171-182]
MiniMax-M3	Acc.	35.3 [33.9-36.7]	59.6 [58.2-60.9]	54.5 [52.9-56.2]	64.5 [63.2-65.9]	33.7 [31.3-36.1]	34.0 [32.2-35.9]	40.4 [38.9-41.8]	64.0 [61.9-66.1]	63.7 [62.2-65.2]	65.0 [62.8-67.3]	64.4 [62.6-66.2]
	Hit	0.0 [n/a]	91.7 [91.7-91.7]	91.2 [91.2-91.2]	94.6 [94.6-94.6]	33.7 [33.7-33.7]	33.7 [33.7-33.7]	89.3 [89.3-89.3]	96.9 [96.5-97.3]	95.7 [95.1-96.4]	94.3 [93.7-94.8]	93.2 [92.7-93.7]
	Tok.	670/113 [670-670/112-113]	5,085/132 [5,085-5,085/131-133]	10,765/160 [10,765-10,765/159-162]	7,111/145 [7,111-7,111/140-149]	6,614/131 [6,614-6,614/126-137]	6,664/137 [6,664-6,664/135-139]	6,684/145 [6,684-6,684/143-148]	5,825/135 [5,806-5,853/134-137]	5,821/136 [5,813-5,830/135-137]	6,513/137 [6,491-6,533/134-139]	6,593/137 [6,569-6,602/134-139]

Note: Acc. = exact-match accuracy; Hit = document hit rate (%); Tok. = mean input/output tokens. Smaller text gives 95% intervals. Zero-shot generation retrieves no context, so its document hit rate is fixed at 0% and an interval is not applicable (n/a). Graph+MM = direct graph-multimodal candidate concatenation. Bold marks the best point estimate within each metric row.

4.2.1 Late interaction is useful but insufficient

ColQwen beats ColPali by 6.4–13.9 percentage points across the five generators and achieves an 89.3% document hit rate, compared with 24.4–33.7% for ColPali. Nevertheless, it remains 18.0–24.1 percentage points behind multimodal re-

trieval. The result separates retrieval from evidence sufficiency: page-level late interaction can locate the linked document while failing to isolate the decisive cell, caption, dose, or endpoint, or while presenting too much page content for reliable generation. Thus ColQwen is a useful retrieval component, not a sufficient end-to-end so-

lution.

4.2.2 Difficulty-stratified patterns

Figure 3 summarizes the ten-run difficulty-stratified distributions for the best-performing Gemma-4-31B. Graph-only retrieval reaches 92.4–93.4% mean accuracy on easy questions but falls to 31.8–35.7% on hard questions. Multimodal retrieval is more balanced, reaching 87.1%, 68.3%, and 44.6% on easy, medium, and hard questions, respectively. Graph-MM 2-Hop matches the hard-question accuracy of multimodal retrieval (44.6%) but does not improve overall accuracy; it also increases mean total tokens from 5,708 to 6,262 and reduces document hit rate from 94.6% to 92.7%. Thus, the figure shows that high document hit rate does not necessarily correspond to answer-sufficient evidence.

4.2.3 Direct fusion rarely improves on its components

Across the repeated panel, direct graph-multimodal fusion ranges from 58.1% to 66.1% accuracy. It exceeds the stronger same-depth component only for Qwen3.6-35B-A3B at both graph depths and MiniMax-M3 at one hop, with gains of 0.1–0.6 percentage points. Direct candidate concatenation therefore does not reliably exploit the two evidence sources and can allow redundant or indirect graph context to displace more precise visual or textual evidence.

4.2.4 Complementarity and unrealized fusion potential

Graph and multimodal retrieval can remain useful even when direct fusion does not improve average accuracy, because the two routes answer different questions correctly. Table 3 shows that both routes answer 43.8–53.2% of questions correctly, while multimodal-only answers account for a further 11.5–16.2% and graph-only answers for 7.9–13.9%. Consequently, ideal selection reaches 69.5–77.7%, compared with 58.1–66.1% for direct concatenation. The largest gap occurs for Qwen3.6-27B with two-hop graph retrieval: 13.5 percentage points, corresponding to approximately 27 additional correct answers per run. The smallest gap occurs for Gemma-4-31B at two hops, but still amounts to 8.4 points, or approximately 17 additional answers. Because ideal selection retrospectively chooses the route that answered each question correctly, these gains are not

directly achievable without a router that can predict the better evidence source before generation. They nevertheless show that all evaluated models contain substantial unrealized complementarity.

4.3 Stage 3: robustness probes and contamination considerations

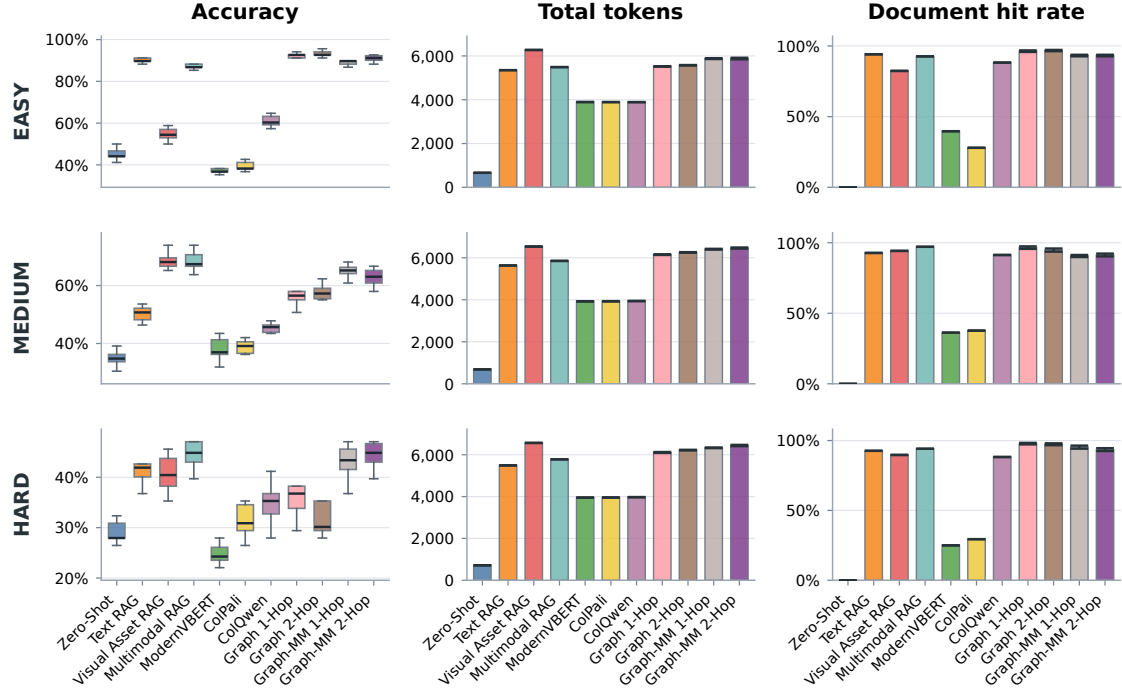
Answer shuffling and context noise change accuracy by at most 1.6 points (Table 4). Paraphrasing is stable for MiniMax-M3 and both Gemma models, but Qwen3.6-27B text retrieval falls from 58.0% to 48.9% ($Z = 2.02$, $p = 0.044$) and Qwen3.6-35B from 55.9% to 42.8% ($Z = 3.13$, $p = 0.002$). Stable zero-shot paraphrasing for both Qwen models localizes this fragility to the tested text-retrieval pipeline rather than the family in every setting.

These perturbations do not diagnose contamination: a drop can reflect prompting, retrieval, or scoring rather than memorization. The Qwen contrast cannot distinguish generator sensitivity from changed retrieval matches or evidence composition. Its immediate value is robustness auditing. Future checkpoint regressions could trigger a separate contamination investigation using meaning-preserving variants plus completion- or overlap-based methods (Yang et al., 2023).

Generated distractors increase accuracy by 12.8–28.7 points ($p < 0.001$ throughout), indicating easier items rather than better pharmacological reasoning. Expert alternatives often encode adjacent doses, endpoints, populations, or domain confusions; generic generation can yield fluent but semantically distant, heterogeneous, or easily eliminated options. This agrees with evidence that surface-valid LLM distractors need not capture misconceptions (Feng et al., 2024), whereas choice-aware optimization improves plausibility and discrimination (Lee et al., 2025).

Holding the expert key fixed while replacing only incorrect options may amplify this effect: the key can remain more specific, evidence-aligned, or stylistically distinct. Text retrieval may expose the exact dose or endpoint needed to reject distant alternatives. These hypotheses require blinded plausibility ratings, style checks, evidence-similarity analysis, and credible-misconception annotation. Regenerated items should not be assumed difficulty equivalent without psychometric validation (Balepur et al., 2025).

Removing options reduces exact match to 0.1–2.3% ($p < 0.001$ throughout). Because response



Note: Rows show easy, medium, and hard questions; columns show exact-match accuracy, total tokens per question, and document hit rate. Boxplots summarize ten run-level aggregates. Graph-MM 1-Hop and Graph-MM 2-Hop denote direct graph-multimodal candidate concatenation. Augmentation labels appear along the bottom row only.

Figure 3: Difficulty-stratified performance of Gemma-4-31B across the 11 retained retrieval conditions.

Table 3: Answer-level complementarity and ideal-selection potential averaged over ten runs.

Model	Both routes correct	Multimodal only	Graph only	Direct fusion	Ideal selection	Potential gain
<i>One-hop graph retrieval</i>						
Gemma-4-26B-A4B	48.9% (n = 100)	11.9% (n = 24)	10.7% (n = 22)	58.3% (n = 120)	71.5% (n = 146)	13.2 pp (n = 26)
Gemma-4-31B	53.2% (n = 109)	13.5% (n = 28)	8.1% (n = 17)	65.8% (n = 135)	74.7% (n = 153)	8.9 pp (n = 18)
Qwen3.6-27B	49.4% (n = 101)	15.8% (n = 32)	12.3% (n = 25)	64.9% (n = 133)	77.5% (n = 159)	12.6 pp (n = 26)
Qwen3.6-35B-A3B	43.8% (n = 90)	12.0% (n = 25)	13.9% (n = 28)	58.2% (n = 119)	69.7% (n = 143)	11.5 pp (n = 24)
MiniMax-M3	51.6% (n = 106)	13.0% (n = 27)	12.4% (n = 26)	65.0% (n = 133)	77.0% (n = 158)	12.0 pp (n = 25)
<i>Two-hop graph retrieval</i>						
Gemma-4-26B-A4B	48.4% (n = 99)	12.3% (n = 25)	11.4% (n = 23)	59.1% (n = 121)	72.2% (n = 148)	13.1 pp (n = 27)
Gemma-4-31B	52.7% (n = 108)	13.9% (n = 28)	7.9% (n = 16)	66.1% (n = 136)	74.5% (n = 153)	8.4 pp (n = 17)
Qwen3.6-27B	48.9% (n = 100)	16.2% (n = 33)	12.6% (n = 26)	64.2% (n = 132)	77.7% (n = 159)	13.5 pp (n = 27)
Qwen3.6-35B-A3B	44.3% (n = 91)	11.5% (n = 24)	13.7% (n = 28)	58.1% (n = 119)	69.5% (n = 142)	11.4 pp (n = 23)
MiniMax-M3	51.6% (n = 106)	12.9% (n = 26)	12.1% (n = 25)	64.4% (n = 132)	76.6% (n = 157)	12.2 pp (n = 25)

Note: Percentages are accompanied by the rounded mean number of correct answers per run (n ; 205 questions). “Both routes correct” denotes questions answered correctly by both multimodal and graph retrieval; the route-only columns identify complementary correct answers. Direct fusion is the implemented candidate-concatenation baseline. Ideal selection is an oracle upper bound that counts a question as correct when either route succeeds. Potential gain is ideal selection minus direct fusion, reported in percentage points (pp) and additional correct answers per run.

format and scoring both change, this demonstrates dependence on option anchors but is not directly comparable to open-ended QA performance.

5 Discussion

The results do not support a uniform advantage for graph retrieval. Retrieval architecture interacts strongly with the generator: graph-only retrieval outperforms multimodal retrieval for Qwen3.6-

Table 4: Robustness analysis for selected models across ten runs.

Model	Baseline	Answer shuffle	Context noise	Paraphrase	Generated distractors	Open ended
<i>Zero-shot generation</i>						
Gemma-4-31B	34.0 [31.9–36.0]	34.8 [32.7–36.9] ($Z=-0.18, p=0.85$)	35.2 [33.2–37.3] ($Z=-0.31, p=0.75$)	36.7 [34.6–38.7] ($Z=-0.71, p=0.48$)	62.7 [60.6–64.8] ($Z=-7.95, p<0.001$)	1.2 [0.7–1.6] ($Z=22.82, p<0.001$)
Gemma-4-26B-A4B	32.4 [30.4–34.4]	33.5 [31.5–35.6] ($Z=0.32, p=0.74$)	32.1 [30.1–34.2] ($Z=0.74, p=0.45$)	33.1 [31.0–35.1] ($Z=0.47, p=0.64$)	56.6 [54.5–58.8] ($Z=-6.03, p<0.001$)	0.7 [0.3–1.0] ($Z=24.69, p<0.001$)
Qwen3.6-27B	33.6 [29.9–37.3]	32.1 [28.5–35.8] ($Z=0.82, p=0.41$)	34.1 [30.3–37.8] ($Z=0.30, p=0.76$)	33.9 [30.2–37.7] ($Z=0.34, p=0.73$)	51.5 [47.6–55.5] ($Z=-4.08, p<0.001$)	1.3 [0.4–2.2] ($Z=14.43, p<0.001$)
Qwen3.6-35B-A3B	31.2 [29.2–33.2]	32.1 [30.1–34.1] ($Z=-0.96, p=0.34$)	31.4 [29.4–33.4] ($Z=-0.76, p=0.45$)	30.8 [28.8–32.8] ($Z=-0.58, p=0.56$)	44.2 [42.1–46.4] ($Z=-4.25, p<0.001$)	1.0 [0.6–1.5] ($Z=20.81, p<0.001$)
MiniMax-M3	36.2 [34.1–38.3]	36.1 [34.0–38.2] ($Z=0.29, p=0.77$)	34.6 [32.6–36.7] ($Z=0.71, p=0.47$)	33.6 [31.6–35.7] ($Z=1.00, p=0.31$)	53.1 [50.9–55.2] ($Z=-4.37, p<0.001$)	2.3 [1.7–3.0] ($Z=21.02, p<0.001$)
<i>Text retrieval</i>						
Gemma-4-31B	58.0 [55.9–60.1]	58.7 [56.6–60.8] ($Z=-0.19, p=0.85$)	56.8 [54.6–58.9] ($Z=0.35, p=0.72$)	56.6 [54.5–58.8] ($Z=0.39, p=0.69$)	75.9 [74.0–77.7] ($Z=-5.58, p<0.001$)	0.3 [0.1–0.5] ($Z=34.78, p<0.001$)
Gemma-4-26B-A4B	54.0 [51.9–56.2]	55.3 [53.1–57.4] ($Z=-0.17, p=0.86$)	54.3 [52.1–56.4] ($Z=0.09, p=0.92$)	51.7 [49.5–53.9] ($Z=0.80, p=0.42$)	70.6 [68.7–72.6] ($Z=-4.75, p<0.001$)	0.1 [0.0–0.2] ($Z=34.33, p<0.001$)
Qwen3.6-27B	58.0 [54.1–61.9]	59.3 [55.4–63.2] ($Z=-0.57, p=0.56$)	57.5 [53.6–61.4] ($Z=-0.12, p=0.90$)	48.9 [45.0–52.9] ($Z=2.02, p=0.044$)	71.7 [68.2–75.3] ($Z=-3.94, p<0.001$)	0.3 [0.0–0.7] ($Z=20.23, p<0.001$)
Qwen3.6-35B-A3B	55.9 [53.7–58.0]	56.3 [54.1–58.4] ($Z=-0.59, p=0.56$)	57.3 [55.2–59.4] ($Z=-0.87, p=0.38$)	42.8 [40.7–44.9] ($Z=3.13, p=0.002$)	69.0 [67.0–71.0] ($Z=-4.34, p<0.001$)	0.2 [0.0–0.4] ($Z=33.82, p<0.001$)
MiniMax-M3	59.9 [57.8–62.0]	58.7 [56.5–60.8] ($Z=0.09, p=0.92$)	58.6 [56.4–60.7] ($Z=0.12, p=0.90$)	58.1 [55.9–60.2] ($Z=0.26, p=0.79$)	72.7 [70.8–74.7] ($Z=-4.16, p<0.001$)	0.1 [0.0–0.2] ($Z=35.76, p<0.001$)

Note: Cells report exact-match accuracy (%) and 95% Agresti–Coull intervals; second lines give two-proportion tests against the corresponding baseline. Baselines have no test. Inference is exploratory because it ignores question pairing and repeated-run clustering.

35B-A3B, whereas multimodal retrieval leads for several other models, and direct fusion rarely improves on the stronger component. Nevertheless, Table 3 shows substantial answer-level complementarity. This finding is consistent with both the promise of joint visual–text retrieval (Chen et al., 2022) and evidence that multimodal retrieval remains vulnerable to tables and rephrasing (Wasserman et al., 2025).

Graph retrieval identifies related sources and entities, but many PK/PD answers require an exact text window, table cell, figure label, dose, or endpoint. Direct concatenation can therefore displace decisive evidence with relevant but insufficient context. Figure 3 likewise distinguishes graph relevance, document hit rate, and answer sufficiency.

Unlike our earlier conversion-focused study (Kocbek et al., 2025), this work evaluates typed MinerU evidence, LightRAG depth, graph–multimodal fusion, and local deployment. The results show that graph augmentation is not consistently effective.

5.1 Future multimodal–graph fusion

In future work, we plan to examine paired item-level tests, inexpensive routing or reciprocal rank fusion (Cormack et al., 2009), semantic or expert scoring of open-ended answers, evidence-level recall, complementarity stratified by difficulty and evidence modality, and question-only re-

trieval with options supplied only to the answering model.

Sparse mixture-of-experts models also merit evaluation on compact systems. DGX Spark provides 273 GB/s memory bandwidth versus 4.8 TB/s for one H200, a 17.6-fold difference (NVIDIA Corporation, 2026a,b). On our H200, dense Gemma-4-31B generated below 10 output tokens/s, while Gemma-4-26B-A4B generated about 50 tokens/s; their best repeated accuracies were 66.6% and 60.8%, respectively. Because model family, active parameters, quantization, and retrieval differ, this is not a controlled hardware comparison. Future work should report accuracy per watt, questions per joule, latency, throughput, and cost under matched settings.

Perturbation trajectories can reveal regressions, but memorization claims require overlap or completion evidence and fresh held-out questions (Yang et al., 2023).

6 Limitations

The single-run screen supports model selection rather than definitive ranking. Ten-run intervals quantify run variability; paired question-cluster intervals remain conditional on the observed runs and do not model source-paper dependence. Robustness tests pool repeated binary responses and may overstate precision. Ideal selection is an oracle bound, not a deployable policy.

The 205-question benchmark is restricted to antimicrobial PK/PD and primarily uses MCQs. Options in retrieval may enable cues or elimination, while option-free exact match may undercount partially correct answers. Generalization requires larger independent benchmarks and evidence-grounded open-ended evaluation. Exact checkpoint revisions should be archived.

Ethical and deployment considerations. The system supports literature inspection, not clinical decision-making. Local deployment may limit document disclosure but does not establish privacy, security, or clinical validity. Users should inspect cited evidence and distinguish retrieved material from generated synthesis; the benchmark contains no patient records.

7 Conclusion

Across 13 local generators, graph retrieval benefits were model- and question-dependent. Multimodal retrieval achieved the highest repeated accuracy; direct fusion rarely improved on the stronger component despite answer-level complementarity. ColQwen outperformed ColPali but remained below multimodal retrieval. These results support learned, source-aware routing or reranking.

Acknowledgments

This research was supported by the EU HE RIA, grant #101159018; the EC/EuroHPC JU and the Slovenian Ministry of ESF through SLAIF, grant #101254461; and the Slovenian Research and Innovation Agency (ARIS), grants #N3-0307 and #GC-0001.

References

Alan Agresti and Brent A. Coull. 1998. [Approximate is better than “exact” for interval estimation of binomial proportions](#). *The American Statistician*, 52(2):119–126.

Nishant Balepur, Rachel Rudinger, and Jordan Boyd-Graber. 2025. [Which of these best describes multiple choice evaluation with LLMs? a\) forced b\) flawed c\) fixable d\) all of the above](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 3394–3418.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. [MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of SIGIR*, pages 758–759.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, et al. 2024. [From local to global: A graph rag approach to query-focused summarization](#). *arXiv preprint arXiv:2404.16130*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Celine Hudelot, and Pierre Colombo. 2025. [Colpali: Efficient document retrieval with vision language models](#). In *Proceedings of the 13th International Conference on Learning Representations*.

Wanyong Feng, Jaewook Lee, Hunter McNichols, Alexander Scarlato, Digory Smith, Simon Woodhead, Nancy Ornelas, and Andrew Lan. 2024. [Exploring automated distractor generation for math multiple-choice questions via large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3067–3082.

Rentong Guo, Xiaofan Luan, Long Xiang, Xiao Yan, Xiaomeng Yi, Jigao Luo, et al. 2022. [Manu: A cloud native vector database management system](#). *Proceedings of the VLDB Endowment*, 15(12):3548–3561.

Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. 2025a. [Rag-anything: All-in-one rag framework](#). *arXiv preprint arXiv:2510.12323*.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025b. [Lightrag: Simple and fast retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10746–10761. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. [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*, pages 39–48.

Primoz Kocbek, Azra Frkatovic-Hodzic, Dora Lalic, Vivian Hui, Gordan Lauc, and Gregor Stiglic. 2025. [Exploration of augmentation strategies in multi-modal retrieval-augmented generation for the biomedical domain: A case study evaluating question answering in glycobiology](#). In *Proceedings of the 2025 IEEE International Conference on Big Data (BigData)*, pages 5156–5165. IEEE.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). In *Proceedings of the 29th ACM Symposium on Operating Systems Principles*, pages 611–626.

- Yooseop Lee, Suin Kim, and Yohan Jo. 2025. [Generating plausible distractors for multiple-choice questions via student choice prediction](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 23669–23692.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems* 33.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Yubo Ma, Jinsong Li, Yuhang Zang, Xiaobao Wu, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2025. [Towards storage-efficient visual document retrieval: An empirical study on reducing patch-level embeddings](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19568–19580.
- Yu A. Malkov and Dmitry A. Yashunin. 2020. [Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.
- NVIDIA Corporation. 2026a. NVIDIA DGX Spark: Hardware overview. <https://docs.nvidia.com/dgx/dgx-spark/hardware.html>. Accessed 2026-08-03.
- NVIDIA Corporation. 2026b. NVIDIA H200 Tensor Core GPU. <https://www.nvidia.com/en-us/data-center/h200/>. Accessed 2026-08-03.
- Pouya Pezeshkpour and Estevam Hruschka. 2024. [Large language models sensitivity to the order of options in multiple-choice questions](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. [Large language models can be easily distracted by irrelevant context](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.
- Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, et al. 2024. [Mineru: An open-source solution for precise document content extraction](#). *arXiv preprint arXiv:2409.18839*.
- Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, et al. 2021. [Milvus: A purpose-built vector data management system](#). In *Proceedings of the 2021 International Conference on Management of Data*, pages 2614–2627. Association for Computing Machinery.
- Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. 2025. [REAL-MM-RAG: A real-world multi-modal retrieval benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 31660–31683.
- Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. 2023. [Rethinking benchmark and contamination for language models with rephrased samples](#). *arXiv preprint arXiv:2311.04850*.