Hybrid Retrieval-Augmented Generation for Chest-Pain Differential Diagnosis: Recall Gains and Calibration Costs

Chinguun Bold[1], Rao Faizan[1], Cam-Hao Hua[2],
Sun Moo Kang[1], Sungyoung Lee[1,*], Seong Tae Kim[1,*]
[1]Kyung Hee University, Yongin, Republic of Korea
[2]The Saigon International University, Ho Chi Minh City, Vietnam
[*]Corresponding authors
chinguun.bold@khu.ac.kr, rao.faizan@khu.ac.kr, huacamhao@siu.edu.vn,
etxkang@khu.ac.kr, sylee0301@gmail.com, st.kim@khu.ac.kr

***Abstract*—Retrieval-Augmented Generation (RAG) is the standard way to ground clinical large language models, but its effect on prediction outside the source dataset has not been studied at case level. We compare three retrieval variants (vector, graph, and hybrid) and a no-retrieval baseline on chest-pain differential diagnosis; all four share one generator, one prompt, and one clinician-authored source dataset, so only the retrieval mechanism varies. Evaluation uses CardioDDxCases, a new benchmark of 165 PubMed Central chest-pain case vignettes with an explicit matched/out-of-distribution partition. Hybrid retrieval improves Top-1 and Top-3 differential recall on covered cases, but on out-of-distribution cases the three retrieval variants commit to a clinician-authored disease at least 87% of the time, while the no-retrieval baseline's top-1 falls outside the dataset on 65% of those cases. Of those 65 cases, 11% name the clinically correct out-of-coverage condition; the remainder are wrong but not in-coverage. In clinical terms, retrieval helps when the patient's true diagnosis is one the dataset already covers, but it becomes risky when it is not: the model keeps naming a familiar cardiac disease instead of stepping outside the list. For a chest-pain triage tool, this means hybrid retrieval can raise recall on common in-scope diagnoses while missing the rarer out-of-scope conditions that most need to reach the clinician. A secondary evaluation on 667 question-answer pairs with the RAGAS suite reproduces the standard RAG metrics and confirms that they do not surface this trade-off. We also show a simple mitigation: a rule that defers to the no-retrieval baseline when it predicts outside the source dataset restores out-of-distribution recognition from 0.01 to 0.65 at no cost to in-coverage recall. Standard RAG metric protocols do not capture the trade-off that this rule targets.**

***Index Terms*—retrieval-augmented generation, knowledge graphs, clinical decision support, differential diagnosis, out-of-distribution detection.**

I. Introduction

Chest pain is a frequent presenting complaint in emergency care. The clinician must build and rank a broad differential within minutes, spanning cardiac, pulmonary, gastrointestinal, musculoskeletal, and psychiatric causes. Choosing correctly matters for survival and resource use, and the information that separates these diagnoses is mostly tabular: pain character, lab values, aggravating factors, and so on.

Large language models read clinical text well and increasingly support clinical tasks from differential diagnosis [1] to medical visual question answering [2], but often get specialised terms wrong without external grounding. Retrieval-Augmented Generation (RAG) [3] is the usual remedy. Vector retrieval is the dominant substrate [4], [5], graph-based retrieval is the main alternative [6]–[9], including in medicine [10], and a hybrid of the two has been studied in the financial domain [11]. None of this prior work measures retrieval's effect on *calibration*: whether the model is willing to predict outside the source dataset when no covered answer fits. In medicine this matters. An overconfident answer on an out-of-coverage case can be worse than no answer.

Recent work moves past fixed retrieval. Self-RAG trains the generator to decide when to retrieve and to critique its own output [12]; clinical systems such as Almanac pair retrieval with curated medical sources [13]; and instruction-tuned medical models such as Med-PaLM reach strong benchmark accuracy with no retrieval at all [14]. We do not benchmark against these systems directly, because each varies the generator, prompt, and corpus together, which would confound the one variable we isolate: the retrieval mechanism over a single fixed clinician-authored source. Our no-retrieval baseline and the abstention rule in §IV-E probe the same adaptive-retrieval question Self-RAG raises, but under a controlled comparison.

We compare three retrieval variants (vector, graph, hybrid) and a no-retrieval baseline on chest-pain differential diagnosis. All four systems share the same generator, prompt, decoding configuration, and clinician-authored source dataset (78 diseases by 16 attributes, with no language-model extraction in either pipeline); only the retrieval mechanism differs. We evaluate on a case-level benchmark with an explicit matched/out-of-distribution partition, and on a secondary RAGAS-style question-answering protocol [15]. Our goal is a controlled evaluation of retrieval behaviour over a small clinician-authored knowledge source, not broad medical coverage.

The main finding is a calibration trade-off. Retrieval im-

979-8-3195-2195-8/26/$31.00

proves recall on covered cases, but on out-of-distribution input every retrieval variant almost always commits to a clinician-authored disease, while the no-retrieval baseline's top-1 falls outside the dataset on 65% of those cases. The over-commitment concentrates on a small set of common cardiac diagnoses. Standard RAG evaluation protocols miss this because they presume the gold answer is in coverage.

This framing matters for reading the LLM-only column. The HybridRAG protocol we build on [11] compares VectorRAG, GraphRAG, and HybridRAG on F/AR/CP/CR without including a no-retrieval column, and this pattern is common in domain-specific RAG method papers: retrieval variants are pitted against each other, and the no-retrieval generator is treated as a side baseline, not as the column that reveals what always retrieving costs. When the source dataset is small and bounded, only the no-retrieval column is willing to predict outside it. That is what makes it the column where the calibration cost of retrieval shows up.

A. Contributions

The novelty is to measure, at case level, what retrieval does to a clinical generator's willingness to predict outside its source dataset, a calibration effect that prior RAG evaluations do not report. Concretely we contribute the following. *(i)* A controlled comparison of vector, graph, and hybrid retrieval against a no-retrieval baseline on chest-pain differential diagnosis, with the generator, prompt, decoding, and source dataset held fixed so that only the retrieval mechanism varies. *(ii) CardioDDxCases*, a benchmark of 165 chest-pain case vignettes from PubMed Central, with verbatim author-stated gold diagnoses, diagnosis-redacted case bodies to prevent answer leakage, an explicit MATCHED/OOD partition, and no language model in the scoring path. *(iii)* Evidence that retrieval improves recall on covered cases but suppresses the model's willingness to predict outside the clinician-authored disease set on out-of-distribution cases. *(iv)* A structural explanation for GraphRAG's top-1 ceiling in terms of entity-linking coverage and diagnosis redaction in case bodies.

II. Methodology

Existing public clinical question-answering datasets [5] probe general medical knowledge but do not pose case-level differential-diagnosis queries against a bounded ontology. We therefore introduce a clinician-authored dataset as the retrieval source, derive a knowledge graph from it, and describe the four retrieval systems we run over both. The evaluation data is described in §III.

A. Clinician-authored dataset

The retrieval source is a clinician-authored dataset structured as a table of 78 chest-pain diseases (rows) by 16 clinical attributes (columns); each non-empty cell encodes how a given disease presents on the attribute. The dataset is obtained by joining two clinician-revised CSVs. A primary source (38 rows, 10 columns) is organised along the OPQRST teaching framework (onset, provocation, quality, region or radiation, severity, timing) together with history and management. A secondary source (79 rows, 12 columns) extends coverage by 41 entries, mostly non-cardiac differentials, and adds workup attributes the primary does not cover (key features, risk factors, diagnostic procedure, test result, lab value, self-management). Joining is on disease name through a curated alias map; on shared attributes, the primary wins.

B. VectorRAG

Each disease row is split into one passage per non-empty attribute cell, so passages fall on predicate boundaries (pain character, diagnostic procedure, lab value, and so on). The four passages with the highest cosine similarity to the query embedding reach the generator. Fig. 1 summarises the indexing path.

C. Knowledge graph construction

Each non-empty cell yields a triplet *(disease, has-predicate, value)* with the predicate read off the column header. A curated alias map handles naming variants; a *variant-of* edge records subtype relations such as Pancoast tumour to lung cancer. Nodes are typed by predicate, so the same surface form under different predicates remains distinct. The construction uses no language-model extraction. Because every non-empty cell maps to a triplet, the graph is dense by construction and needs no link-prediction step to recover missing edges [16].

D. GraphRAG

The entity linker matches the query against node labels, prefers the longest match, and breaks ties by node degree. If no disease label matches, the linker falls back to up to five attribute-value nodes mentioned in the query, surfacing diseases that share those attributes. From the seeds we walk one hop, cap the result at fifty triplets, and serialise them as (head, relation, tail) lines. When nothing matches, the subgraph is empty and the generator is told so. Fig. 2 shows the graph construction.

E. HybridRAG

HybridRAG concatenates the vector context with the graph context, in that order, and feeds the joined context to the same generator. Vector contributes similarity-based recall over the

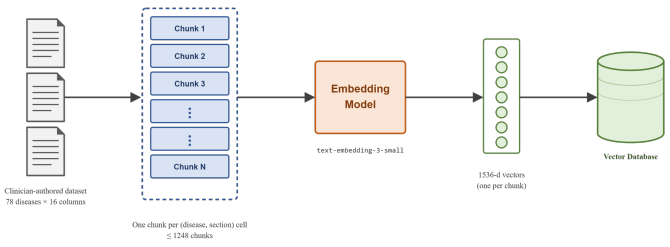

Fig. 1. VectorRAG indexing. Each non-empty (disease, attribute-column) cell of the dataset yields one passage; passages are embedded with `text-embedding-3-small` and stored in a vector database.

dataset's text; graph contributes relational evidence around the linked entities. Fig. 3 shows the query-time pipeline end to end.

III. Experimental Setup

With the four systems defined, we now describe how we evaluate them. Decoding is deterministic (temperature zero). Per-system configurations appear in Tables I and II; all retrieval hyperparameters were fixed before evaluation and were not tuned on CardioDDxCases. For HybridRAG, the vector context precedes the graph context in the concatenation; items in the graph block therefore appear later in the joined context than items in the vector block, which interacts with the RAGAS context-precision judge as we revisit in §IV-D.

A. Benchmark construction

1) CardioDDxCases dataset: CardioDDxCases is designed as a safety test, not just another chest-pain dataset. The MATCHED/OOD split and the redaction of the final diagnosis from each case body target one specific failure mode: a retrieval system that confidently names a covered disease on a case that is in fact out of coverage. Placing MATCHED and OOD cases under the same protocol lets us measure in-coverage recall and out-of-coverage over-commitment together, using the same generator and the same prompt. Each case is labelled MATCHED (the verbatim author-stated diagnosis maps cleanly into the dataset) or OOD (no plausible counterpart).

Cases are drawn from the PubMed Central Open Access subset [17] through a three-step deterministic pipeline. *(1) Probe.* NCBI E-Utilities queries retrieve chest-pain case reports; results are filtered to the `case-report` article type and to the CC BY 4.0 license, so every shipped record is redistributable. *(2) Curate.* For each case body a single language-model call extracts four fields: chief complaint, presentation summary, verbatim final diagnosis, and three to seven key diagnostic features. The extraction prompt redacts the final diagnosis, the confirming imaging, and the treatment outcome from the presentation summary, so the eval-time input never contains the answer. A verifier then rejects any extraction whose final diagnosis still appears in the summary,

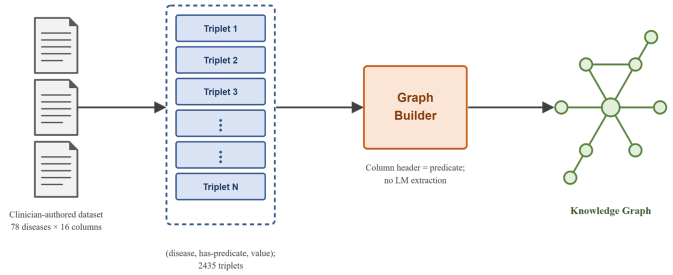

Fig. 2. GraphRAG construction. Column headers supply the predicate vocabulary, so triplets are obtained by structural decomposition without a language model in the loop; the resulting Knowledge Graph holds 1688 nodes and 2435 edges.

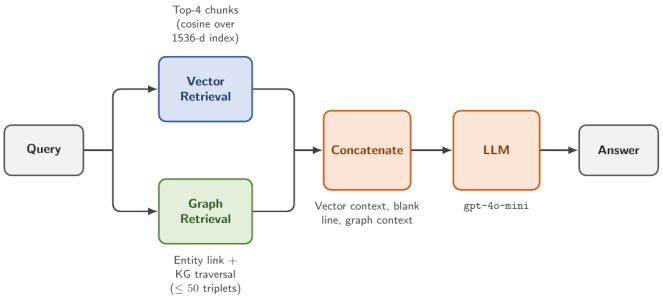

Fig. 3. HybridRAG at query time. The query is routed in parallel to both retrievers; their outputs are concatenated (vector context, blank line, graph context) into a single prompt and passed to the generator.

and rejected cases are logged rather than silently dropped. *(3) Match.* The MATCHED/OOD label is assigned with no language model, by `difflib` sequence similarity between the verbatim diagnosis and the 78 canonical disease names together with a curated alias and subtype map. A case is MATCHED when the best similarity is at least 0.55 and OOD otherwise. The matched surface form and its similarity score are stored on every record, so the split is auditable and byte-reproducible. The language model takes part only in field extraction: gold answers are the case authors' own conclusions, and no model takes part in scoring.

2) Question/answer dataset: For the secondary track we use 667 question/answer pairs from three independently sourced subsets, none generated by the system under evaluation: *Extractive* (312 items, template-instantiated over *(disease, predicate, value)* triplets from the dataset), *MedQuAD* (350 items [18], filtered to focus terms on the chest-pain dataset), and *K-QA* (5 items [19], with sentence-level must-have annotations useful for CR).

B. Evaluation metrics

We report two families. **Top-K accuracy** is the indicator that the gold disease g_c appears within the first K positions of the predicted differential $\hat{y}_c$, averaged over the MATCHED subset $\mathcal{M}$:

$$\text{Top-}K \;=\; \frac{1}{|\mathcal{M}|}\sum_{c\in\mathcal{M}} \mathbf{1}\big[\,g_c \in \text{top}_K(\hat{y}_c)\,\big], \qquad K \in \{1,3,5\}.$$

OOD-Rec is the fraction of OOD cases on which the system's top-1 either is empty or does not fuzzy-match any of the 78 clinician-authored disease names. Both a refusal and a clinically correct out-of-dataset name (e.g., *Takotsubo cardiomyopathy*) contribute; the metric captures *dataset anchoring*. The converse, $1 -$ OOD-Rec, is the fraction of OOD cases where the system commits to a clinician-authored disease that does not fit.[1]

The second family is the four-metric RAGAS suite [15]: Faithfulness (F, fraction of generated statements entailed by

[1]We report OOD-Rec in the positive form so that higher values correspond to better behaviour on out-of-coverage input, mirroring the Top-K convention.

the context), Answer Relevance (AR, similarity between the original question and questions an LLM judge reconstructs from the answer), Context Precision (CP, precision-at-K over a per-rank relevance indicator), and Context Recall (CR, fraction of gold-answer sentences supported by the context). Definitions follow [15].

C. Environment configurations

Tables I and II summarise the per-system configurations.

IV. Results

Results are presented in two parts: case-level findings on CardioDDxCases (§IV-A–IV-C) and the RAGAS Q&A track (§IV-D). Top-K accuracy is averaged over the 65 MATCHED cases and OOD-Rec over the 100 OOD cases. The LLM-only column is the same generator without retrieved context and without any instruction to anchor on the clinician-authored disease set.

A. Recall on MATCHED cases

Table III shows that retrieval improves Top-K recall on MATCHED cases, especially at deeper ranks.

All four systems run on the same 65 MATCHED cases, so between-system differences can be assessed with paired McNemar 95% intervals. The Top-3 and Top-5 gaps between HybridRAG and VectorRAG (both $+0.077$, ± 0.065) and the Top-1 gap between HybridRAG and LLM-only ($+0.169$, ± 0.132) exceed their intervals and are statistically separated at this N. The Top-1 gap between HybridRAG and VectorRAG ($+0.031$, ± 0.073) sits inside its interval and should be read as directional. At Top-3 and Top-5, HybridRAG and LLM-only are tied (both gaps inside their paired intervals), so retrieval's main statistical advantage over no-retrieval at this benchmark size is at Top-1.

TABLE I
VectorRAG configuration.

LLM	gpt-4o-mini
LLM Temperature	0
Embedding Model	text-embedding-3-small
Embedding Dimension	1536
Maximum Output Tokens	512
Number of Indexed Passages	979
Number of Context Retrieved	4

TABLE II
GraphRAG configuration.

LLM	gpt-4o-mini
LLM Temperature	0
KG Manipulation	NetworkX
Number of Triplets	2435
Number of nodes	1688
Number of edges	2435
BFS Depth	1
Triplet Cap per Query	50
Maximum Output Tokens	512

TABLE III
Primary results on CardioDDxCases. Top-K averaged over the 65 MATCHED cases; OOD-Rec averaged over the 100 OOD cases. Retrieval improves Top-K recall on MATCHED cases but strongly anchors predictions to the clinician-authored disease set on OOD cases.

	Top-1	Top-3	Top-5	OOD-Rec
VectorRAG	0.48	0.54	0.54	0.04
GraphRAG	0.42	0.48	0.48	0.13
HybridRAG	**0.51**	**0.62**	**0.62**	0.01
LLM-only	0.34	0.54	0.58	**0.65**

HybridRAG has the highest Top-1 score, though the margin over VectorRAG is small on a 65-case benchmark. The larger difference appears at Top-3 and Top-5, where HybridRAG reaches 0.62 against 0.54 for VectorRAG. The graph component helps recover diseases that exist in the dataset but do not appear among VectorRAG's nearest dense neighbours. VectorRAG itself plateaus after Top-3: at this corpus size, dense retrieval finds the correct answer in the first three ranks or not at all. The LLM-only baseline overtakes GraphRAG at $K \geq 3$: an under-linked graph context can be worse than no retrieved context at deeper ranks. GraphRAG's lower Top-1 has two structural causes, unpacked in §IV-C. These recall gains come with the dataset-anchoring effects examined next.

B. Dataset anchoring on OOD cases

While MATCHED measures recall on covered cases, the OOD subset measures what each system does when no covered answer fits. We use *dataset anchoring* to describe cases where the generator continues committing to diagnoses inside the fixed 78-disease clinician-authored set even when no covered disease fits the case. This is the paper's main finding. On 65 of the 100 OOD cases the LLM-only baseline's top-1 lies outside the clinician-authored disease set. These cases split into:

- 1 explicit refusal ([ ]);
- 7 clinically correct out-of-dataset diagnoses (top-1 fuzzy-matches the author-stated gold at ≥ 0.70, e.g. *Takotsubo cardiomyopathy*, *Kawasaki disease*, *Thyrotoxic periodic paralysis*);
- 16 partial paraphrases of the gold;
- 41 unrelated out-of-dataset predictions.

The three retrieval variants collapse on this dimension: VectorRAG's top-1 stays inside the clinician-authored set on 96% of OOD cases, HybridRAG on 99%, and GraphRAG on 87% (the last only because the substring linker fails to seed, leaving the generator without context). **In other words, retrieval improves differential recall at the cost of dataset anchoring on uncovered cases.** Two features of the retrieval setup drive this. The prompt presents the clinician-authored disease set as the answer space, and the retrieved context only ever contains in-set diseases, so the generator is never shown an out-of-set option. When no covered disease fits, the model picks the nearest in-set name rather than stepping outside the list.

Standard RAG evaluation protocols do not capture this trade-off.

The over-commitment is highly concentrated. Across the three retrieval systems, the five most common Top-1 picks account for 45–47% of all OOD predictions, dominated by myocardial infarction, unstable angina, and pulmonary embolism, the dataset's most common cardiac entries. When the LLM-only baseline does commit to a clinician-authored disease (35 cases), it spreads its predictions more evenly: its five most common picks account for only 34% of those cases. Retrieval does not only bind the generator to the clinician-authored disease set; it also nudges it toward the same cardiac diagnoses that emergency clinicians are already prone to over-call. Fig. 4 shows the retrieval systems repeatedly collapsing onto the same handful of common cardiac diagnoses.

C. Structural causes of GraphRAG's Top-1 deficit

We next look at why GraphRAG does worse at Top-1, given HybridRAG's gains at deeper ranks.

Seeding ceiling. GraphRAG Top-1 is only 0.12 on unseeded cases, but rises to 0.74 when the gold disease appears among the linked seeds. The entity linker (a word-boundary matcher over disease labels and a curated map of clinically-equivalent surface forms) fires on every query, but the gold appears among its seeds on only 31 of 65 MATCHED cases (48%). HybridRAG inherits the same ceiling because its graph block is fed from the same linker; its gains over VectorRAG at $K \geq 3$ come from cases where the gold disease was seeded alongside the vector neighbours.

Diagnosis redaction. Of the 34 unseeded cases, only 5 contain any gold-pointing surface form in the case body; the remaining 29 have been diagnosis-redacted by construction (the curation strips the final diagnosis from the presentation summary to prevent answer leakage). The 48% seeding ceiling therefore reflects both linker policy and diagnosis redaction. Redaction is the larger factor: even a perfect surface-form linker would raise seeding only to about 56%.

Structured confusions. Errors above the ceiling are clinically structured rather than random: the most stable confusion across all three retrieval systems is pericarditis to pulmonary embolism, a pleuritic-pain crossover the dataset does not resolve on its current predicates. Seven of the twenty unique gold diseases on MATCHED are missed at Top-1 by every retrieval system on every case: a long tail concentrated on entries either rarer in case reports or unrecoverable from their redacted bodies.

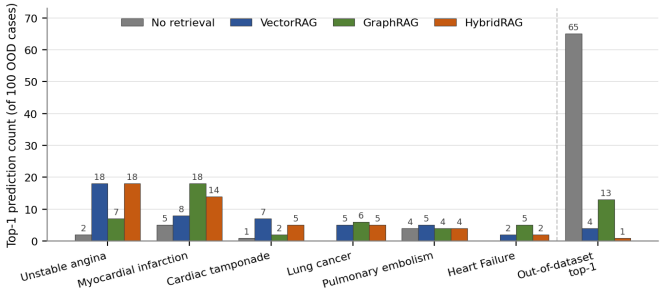

Fig. 4. Top-1 prediction counts on the 100 OOD cases. The no-retrieval baseline frequently predicts outside the clinician-authored disease set, while the retrieval systems concentrate on a small set of common cardiac diagnoses.

D. RAGAS metrics on the Q&A track

Having focused on case-level behaviour, we now turn to the standard RAG evaluation metrics. For comparability with prior RAG work we report the four-metric RAGAS suite on the 667-question Q&A set (Table IV); these scores do not capture the recall/calibration trade-off visible on the primary benchmark. Two artefacts of the suite explain why. Context Precision tracks concatenation order: items in the trailing graph block of HybridRAG's joined context attract more false-negative judgements than items in the leading vector block, even when the same content appears in both. Faithfulness rewards refusals: when the linker fails to seed, the generator returns "insufficient context", which has no statements to verify and scores $F{=}1$ vacuously.

E. A calibration lever: LLM-only abstention

The LLM-only column doubles as a cheap out-of-coverage probe. We test a simple composite rule: abstain when LLM-only's top-1 falls outside the 78-disease set, and otherwise return HybridRAG's top-1. Table V reports the result. The rule recovers OOD-Rec from 0.01 to 0.65 while leaving MATCHED Top-K unchanged: it abstains on none of the 65 MATCHED cases and on 65 of the 100 OOD cases. The clean separation partly reflects benchmark construction, since MATCHED is defined by the same coverage check the probe applies. We report this as a first calibration lever, not a complete solution.

TABLE IV
RAGAS METRICS ON THE 667-QUESTION Q&A SET. F: FAITHFULNESS; AR: ANSWER RELEVANCE; CP: CONTEXT PRECISION; CR: CONTEXT RECALL.

	VectorRAG	GraphRAG	HybridRAG
F	0.91	**0.96**	0.92
AR	0.80	0.80	0.80
CP	**0.68**	0.18	0.39
CR	0.53	0.51	**0.54**

TABLE V
LLM-ONLY ABSTENTION RULE ON CARDIODDXCASES. ABSTAINING WHEN THE NO-RETRIEVAL TOP-1 FALLS OUTSIDE THE 78-DISEASE SET, AND RETURNING HYBRIDRAG'S TOP-1 OTHERWISE, RECOVERS OOD-REC TO THE LLM-ONLY LEVEL AT NO COST TO MATCHED TOP-K.

	Top-1	Top-3	Top-5	OOD-Rec
HybridRAG	0.51	0.62	0.62	0.01
LLM-only	0.34	0.54	0.58	0.65
HybridRAG + abstention	**0.51**	**0.62**	**0.62**	**0.65**

V. Conclusion and Future Directions

We compared three retrieval variants (vector, graph, hybrid) and a no-retrieval baseline on chest-pain differential diagnosis over a single clinician-authored dataset. The headline finding is a calibration trade-off: retrieval improves recall on covered cases but anchors predictions to the clinician-authored disease set on out-of-distribution cases, concentrating on a small set of common cardiac diagnoses that emergency clinicians are already prone to over-call. The construction recipe and the protocol should transfer to any RAG system whose source is a small clinician-authored dataset and whose output is a ranked differential under case-level inputs.

Several limitations apply. The benchmark is restricted to chest pain and the source dataset to 78 diseases by 16 attributes, so coverage and attribute granularity are narrower than a deployed clinical decision-support system would require. The 165 case vignettes are retrospective and literature-derived from PubMed Central, which selects for unusual or publishable presentations rather than the case mix an emergency department actually sees. MATCHED $N{=}65$ is small: as the paired McNemar intervals in §IV-A show, HybridRAG's gains over VectorRAG at Top-3 and Top-5 and over LLM-only at Top-1 are statistically separated at this N, while its Top-1 gap over VectorRAG and its Top-3 and Top-5 gaps over LLM-only are not, and should be read as directional. Because every retrieval system also carries the disease-set framing in its prompt, our design does not separate prompt-induced from context-induced anchoring. Nothing reported here is a deployable system: the trade-off needs prospective validation on broader patient cohorts and on settings beyond chest pain before any clinical use.

The LLM-only abstention rule in §IV-E is a first calibration lever: it recovers OOD-Rec to the no-retrieval level at no cost to MATCHED Top-K. Two follow-ups remain. A finding-to-disease linker that seeds from the symptom and lab predicates actually present in the redacted query would address the seeding ceiling directly. Beyond that rule, a learned calibration layer could exploit signals such as retrieval-score margins or three-way disagreement among the retrieval variants. Longer-term work includes extending CardioDDxCases beyond chest pain and formalising metric variants that bound the two RAGAS artefacts.

Acknowledgment

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (Explainable Logical Reasoning for Medical Knowledge Generation) under Grant RS-2022-II220078; in part by the IITP-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) under Grant IITP-2026-RS-2023-00259004; and in part by the IITP-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT) under Grant IITP-2026-RS-2020-II201489.

References

[1] D. McDuff *et al.*, "Towards accurate differential diagnosis with large language models," *Nature*, vol. 642, no. 8067, pp. 451–457, 2025.

[2] R. Faizan, S. Lee, and S. T. Kim, "ChestVQATrans: A transformer-based approach for medical VQA in radiology," in *Proc. 8th Artif. Intell. Cloud Comput. Conf. (AICCC)*, 2026, pp. 169–173.

[3] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. NeurIPS*, 2020.

[4] V. Karpukhin *et al.*, "Dense passage retrieval for open-domain question answering," in *Proc. EMNLP*, 2020.

[5] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," in *Findings of ACL*, 2024, pp. 6233–6251.

[6] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 3580–3599, 2024.

[7] D. Edge *et al.*, "From local to global: A graph RAG approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.

[8] B. Peng *et al.*, "Graph retrieval-augmented generation: A survey," *ACM Trans. Inf. Syst.*, vol. 44, no. 2, pp. 1–52, 2026.

[9] H. Han *et al.*, "Retrieval-augmented generation with graphs (GraphRAG)," *arXiv preprint arXiv:2501.00309*, 2025.

[10] J. Wu *et al.*, "Medical Graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation," in *Proc. ACL*, 2025, pp. 28443–28467.

[11] B. Sarmah, D. Mehta, B. Hall, R. Rao, S. Patel, and S. Pasquali, "HybridRAG: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction," in *Proc. 5th ACM Int. Conf. AI in Finance (ICAIF)*, 2024, pp. 608–616.

[12] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[13] C. Zakka *et al.*, "Almanac: Retrieval-augmented language models for clinical medicine," *NEJM AI*, vol. 1, no. 2, 2024.

[14] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[15] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval-augmented generation," in *Proc. EACL (System Demonstrations)*, 2024.

[16] T. D. Nguyen *et al.*, "Unified link prediction modeling for enhanced knowledge graph completion task," *Expert Syst. Appl.*, vol. 279, art. no. 127559, 2025.

[17] National Library of Medicine, "PubMed Central Open Access Subset," https://pmc.ncbi.nlm.nih.gov/tools/openftlist/, 2024 (accessed May 2026).

[18] A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC Bioinformatics*, vol. 20, art. no. 511, 2019.

[19] I. Manes, N. Ronn, D. Cohen, R. I. Ber, Z. Horowitz-Kugler, and G. Stanovsky, "K-QA: A real-world medical Q&A benchmark," in *Proc. 23rd Workshop Biomed. Natural Lang. Process. (BioNLP)*, 2024, pp. 277–294.