

Towards guideline-grounded clinical decision-making: a multi-scale study of retrieval augmented generation based large language model agents

Romain André

`Romain.Andre@radboudumc.nl`

Radboud University Medical Center <https://orcid.org/0009-0008-9919-0030>

John Hermans

Radboud University Medical Center <https://orcid.org/0000-0001-9207-0548>

J.P. Sedelaar

Radboud University Medical Center

Henkjan Huisman

Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen

<https://orcid.org/0000-0001-6753-3221>

Article

Keywords: AI-Agents, LLM, RAG, Abdominal, Acute, Clinical decision-making, Guidelines

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11226907/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: **Yes** there is potential Competing Interest. R.A. receives research support through the HealthyAI project (project no. KICH3.LTP.20.006), which is partly financed by the Dutch Research Council (NWO) with co-funding from Siemens Healthineers. H.H. has received research funding from Siemens Healthineers and Canon Medical Systems. J.J.H. and J.P.M.S. declare no competing interests.

Towards guideline-grounded clinical decision-making: a multi-scale study of retrieval augmented generation based large language model agents

Romain André^{1,*}, John J. Hermans², J. P. Michiel Sedelaar³ & Henkjan Huisman¹

¹Diagnostic Image Analysis Group, Department of Medical Imaging, Radboud University Medical Center, Nijmegen, the Netherlands.

²Department of Medical Imaging, Radboud University Medical Center, Nijmegen, the Netherlands.

³Department of Urology, Radboud University Medical Center, Nijmegen, the Netherlands.

*Correspondence to: Romain André, romain.andre@radboudumc.nl

Large language models (LLMs) show promise for clinical decision support but are limited by opaque knowledge and inconsistent decision-making. Retrieval-Augmented Generation (RAG) mitigates these issues by injecting verifiable passages from a curated corpus at inference time. Using 2,400 real patient pathways from the MIMIC-CDM dataset, we evaluated RAG configurations over information gathering, diagnostic accuracy, treatment recommendations and guideline retrieval, identifying key factors for integrating RAG into clinical AI-agents. Across eight general and medical LLMs spanning 1B to 70B parameters, RAG consistently increased mean diagnostic accuracy from 55.64% to 64.29%, an 8.65-points gain ($P=0.0039$). In a matched three-backbone analysis, RAG improved physical-examination, imaging and laboratory clinician test-ordering alignment by 5.17 points and treatment alignment by 3.97 points, while surfacing verifiable guideline citations for clinicians. Benefits depended on retrieval design, with autonomous query rewriting producing the clearest effect on retrieval fidelity and downstream performance. These findings demonstrate that guideline-grounded RAG can improve both the outcomes and clinical-decision pathways of LLM-agents while increasing their traceability and transparency.

Keywords: AI-Agents, LLM, RAG, Abdominal, Acute, Clinical decision-making, Guidelines.

Main

Advancements in large language models (LLMs) have highlighted their potential in healthcare by enabling rapid aggregation and synthesis of large quantities of data such as clinical notes, laboratory reports, and imaging results across increasingly complex patient records¹⁻³. However, clinical usefulness requires more than fluent generation or performance on medical examinations. LLM outputs continue to hold critical limitations, including hallucinations (generating incorrect or fabricated information), lack of transparency from their internal reasoning processes, unknown sources of knowledge, as well as varying with the formulation or ordering of the information provided.

In a sequential diagnostic pathway, these limitations can undermine reliability and trustworthiness, affecting not only the final diagnosis but also which examinations are requested, which findings are ignored and how early the model commits to a diagnosis^{4,5}.

We hypothesize that Retrieval-Augmented Generation (RAG) can mitigate part of these challenges by grounding the model in external, authoritative sources. With RAG, an information-retrieval component selects contextually relevant text passages from a curated knowledge base, which are then supplied to the LLM at inference time⁶. This design appears ideal for clinical deployment; the knowledge can be updated and localized, using hospital-specific guidelines and expert-authored sources, without the cost of retraining or fine-tuning the backbone model. It also exposes the model, and ultimately the clinician, to citable, trusted guidelines supporting a recommendation, enhancing transparency^{6,7}. Yet, RAG is not plug-and-play. Performance is highly sensitive to corpus preparation, chunking, embedding, query formulation, candidate retrieval, and how evidence is presented to the LLM^{8,9}. Moreover, guidelines themselves are often conditional and heterogeneous across evidence levels, organizations, and revisions, so retrieval may surface competing recommendations rather than resolve uncertainty¹⁰⁻¹².

Most evaluations of medical RAG have focused on single-turn question answering, licensing-style questions or narrowly specified clinical scenarios, which do not accurately capture the complexity and iterative nature of clinical decision-making^{1-3,13,14}. The MIMIC-IV-Ext Clinical Decision-Making benchmark (MIMIC-CDM) addresses this by presenting the LLM-Agent with only the patient’s medical history, and requiring it to autonomously request physical examinations, laboratory tests and imaging modalities. Iteratively integrating new findings, the agent stops once it judges sufficient evidence to deliver a diagnosis and care plan. The benchmark comprises 2,400 emergency-department presentations for appendicitis, cholecystitis, diverticulitis or pancreatitis derived from the MIMIC-IV dataset^{15,16}. The original evaluation showed that LLMs frequently made incomplete diagnostic workups and performed worse when required to gather information themselves, providing a suitable setting in which to test whether external guideline evidence changes agent behavior.

In this work, we extend MIMIC-CDM with a guideline-grounded retrieval loop and evaluate eight locally deployable LLM backbones spanning 1B to 70B parameters, and spanning general domain and medical fine-tuning. We first test whether RAG improves diagnostic accuracy across model scales and medical specialization. We then examine whether retrieval changes the diagnostic process, using clinician test-ordering alignment, treatment-ordering, and guidelines retrieval accuracy as a process-level outcome. Finally, through corpus comparisons and component ablations, we determine which elements of retrieval are responsible for any benefit. Our central finding is that guideline retrieval improves iterative clinical decision-making, but that the magnitude and consistency of this improvement depend on structured retrieval, most notably task-focused autonomous query rewriting, rather than on the addition of retrieved text alone.

Results

Adding retrieval to MIMIC-CDM enables evaluation of evidence-grounded clinical agents

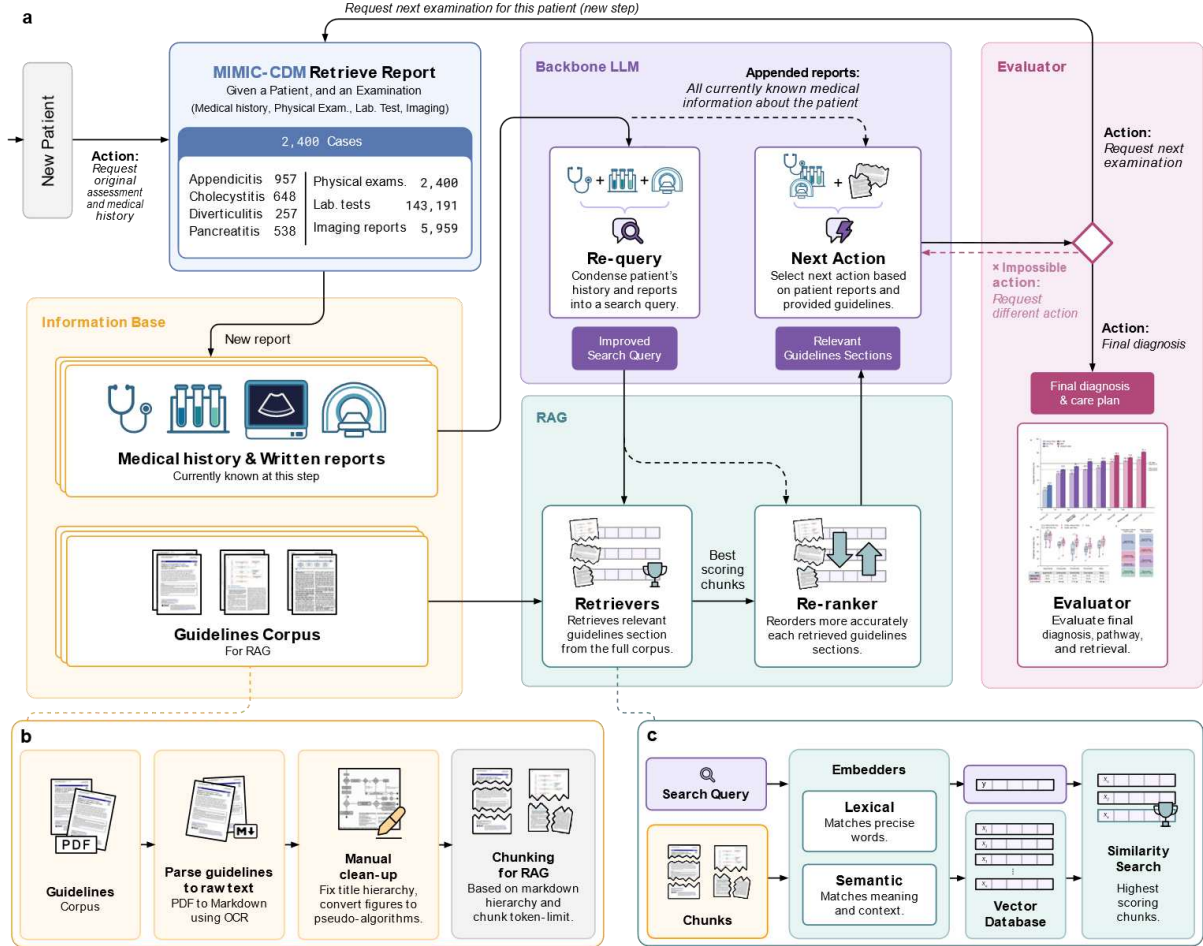


Fig. 1 | Guideline-grounded extension of MIMIC-CDM for iterative clinical decision-making. **a**, Overview of the RAG-augmented MIMIC-CDM framework. The original 2,400 patient trajectories, clinical actions and evaluator are retained and extended with a per-step retrieval loop over a curated clinical-guideline corpus. At each agentic decision step, the accumulated patient history and known reports are autonomously condensed into a focused search query. Relevant guideline sections are retrieved and reranked, then appended to the backbone LLM context together with the patient information before ordering of the next examination or, once sufficient information has been gathered, generation of the final diagnosis and care plan. The evaluator assesses whether next action is valid, if the next action is ordering the next test, it will return requested patient information and advance the case. If the action is a final diagnosis, the evaluator terminates this session and saves this patient's complete decision pathway, allowing performance evaluation on multiple metrics processing. If the action is not valid, the evaluator will request the agent to suggest an alternative action. Disabling the guideline retrieval components yields the matched non-RAG comparator. **b**, Guideline-corpus preparation. Source guideline documents are converted from PDF to machine-readable text, manually cleaned to preserve document structure and remove conversion artefacts, and divided into retrieval chunks according to the document hierarchy and chunk-size limit. **c**, Retrieval and indexing workflow. Guideline chunks are indexed for complementary lexical and semantic retrieval, and the search query is matched against the indexed corpus to identify the highest-scoring candidate chunks for subsequent reranking and inclusion in the LLM context. LLM, large language model; RAG, retrieval-augmented generation.

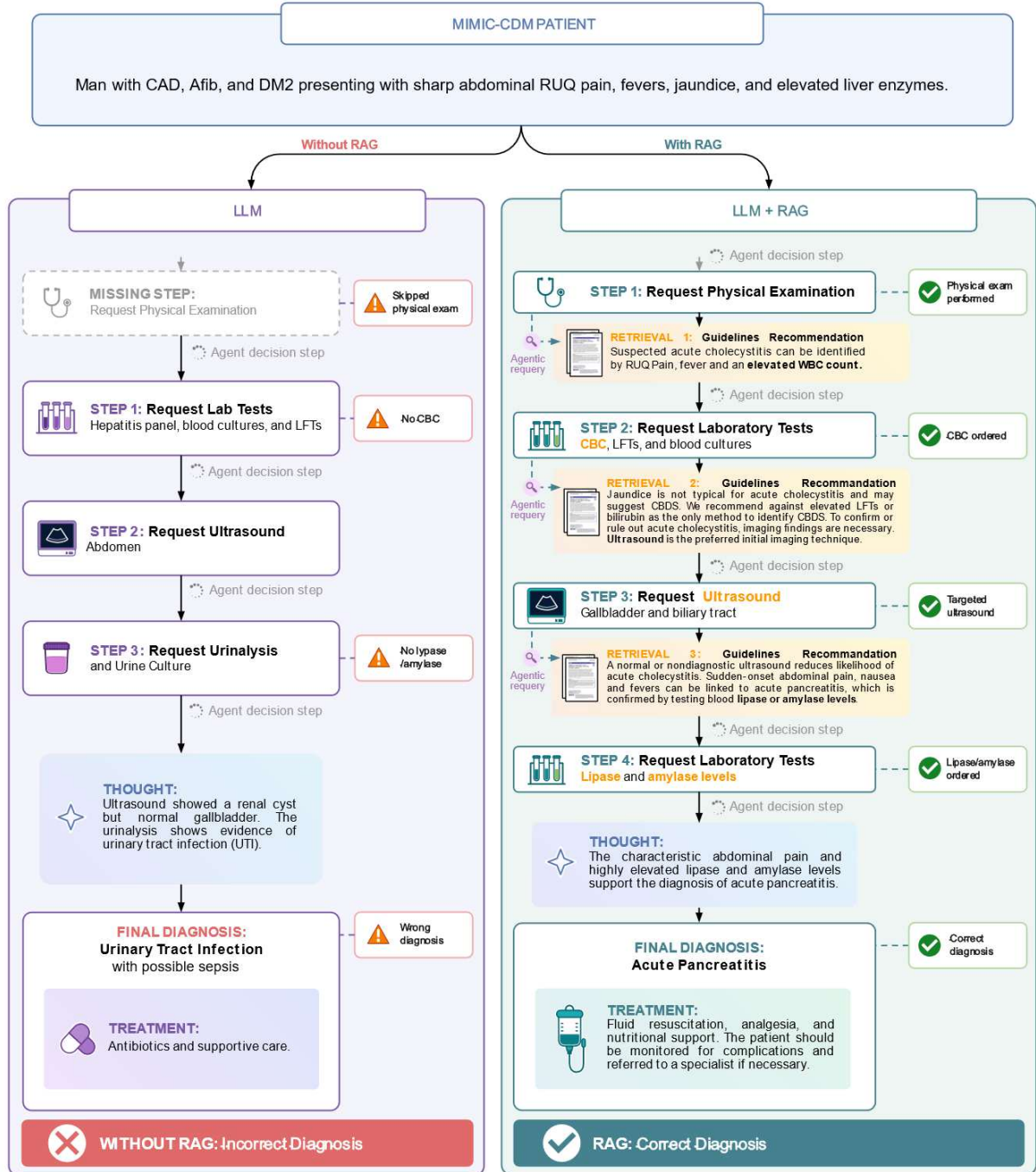


Fig. 2 | Case-level comparison of the same backbone model with and without RAG. This illustrative pathway shows how retrieval can alter both the diagnostic work-up and the final diagnosis. In the non-RAG setting, the model skips key tests and converges on an incorrect diagnosis. Whereas the same backbone model for the same seed and same patient, but equipped with RAG, retrieves guideline passages at each decision step, grounding its tests ordering into a more informed reasoning, and reaches the correct final diagnosis and care plan. This figure represents an actual case from an acute pancreatitis patient taken from the dataset, that we have then manually anonymized. In this comparison, the backbone model is Llama3.1-70B-Instruct^{17,18} quantized in 4 bits-per-weight, with and without RAG, with matched inference settings and ran with seed 2023.

We extended the existing MIMIC-IV-Ext Clinical Decision-Making benchmark (MIMIC-CDM)^{15,16} framework with a guideline-retrieval layer while preserving its 2,400 emergency-department acute abdominal pain patient trajectories, available clinical actions, report-return logic and reference diagnoses which are one of appendicitis, cholecystitis, diverticulitis or pancreatitis¹⁵ (Fig. 1). Each case starts with the patient’s history of present illness. The LLM-Agent

then autonomously requests a physical examination, laboratory tests or imaging study, receives the corresponding patient report and iteratively integrates these findings until it delivers a final diagnosis and care plan. In the retrieval-augmented generation (RAG) condition, the evolving patient state is supplemented with relevant guideline evidence before each decision step. The matched non-RAG condition uses the same backbone model, patient information and decision environment, but does not receive retrieved passages.

Unlike diagnostic evaluations in which all findings are provided simultaneously, MIMIC-CDM requires the agent to decide which clinical information to obtain, in what sequence and when to stop gathering evidence. The evaluator retained the benchmark’s inherited action semantics: available requests returned the corresponding report, whereas unavailable or invalid requests were handled using the predefined MIMIC-CDM behavior. Retrieval therefore changed only the external evidence available to the agent, not the underlying patient information or the execution of clinical actions. In the principal configuration, the first action was generated from the history alone, and retrieval was activated after the first returned clinical observation.

The retrieval knowledge base contained clinical guidance documents covering the four target pathologies and, in selected experiments, additional ambiguous differential diagnoses for acute abdominal pain. Source documents were converted into machine-readable text, hand corrected for document-conversion and recognition artefacts, and divided into retrievable snippets while preserving source and section metadata. In the guidelines, clinically informative figures or diagnostic algorithms were transcribed into machine-readable text when required.

At each decision step, the accumulated patient history and returned reports were rewritten into a concise, task-focused query. A hybrid lexical and dense retrieval system then generated candidate passages, which were reranked according to relevance and assembled into an evidence block before the agent selected its next action. Retrieval was repeated after each new observation, allowing the evidence to evolve with the clinical state. Complete implementation details are provided in Methods and prompts in Extended Data.

This framework allowed us to assess whether retrieval changed not only the final diagnosis, but also the sequence of decisions leading to it. Fig. 2 presents an illustrative pancreatitis case evaluated using the same backbone, inference settings and decoding seed. Without RAG, the model omitted the physical examination, did not request pancreatic-enzyme testing and concluded that the patient had a urinary tract infection with possible sepsis. With RAG, the model first requested the physical examination and then pursued a targeted laboratory and imaging workup. After a nondiagnostic biliary ultrasound, the newly retrieved evidence shifted toward pancreatitis-relevant guidance, prompting measurement of lipase and amylase and leading the model to the reference diagnosis. Although this single case is descriptive and was not used for statistical inference, it illustrates how per-step retrieval can redirect an evolving diagnostic pathway rather than only alter the final response.

This framework therefore enabled evaluation of both final diagnostic performance and two intermediate processes: the clinical information selected by the agent and the pathology relevance of the evidence retrieved during that process.

RAG improves diagnostic accuracy across scales and domains

Table 1 | Overview of evaluated large language models.

Model	Size	Specialization	Notes	Training dataset (as publicly described)
Llama3.2-1B-Instruct ¹⁹	1B	General instruct	Open-weight, Q-4.0bpw	“A new mix of publicly available online data”; up to 9T pretraining tokens; cutoff December 2023.
Mistral-7B-Instruct-v0.3 ^{20,21}	7B	General instruct	Open-weight, Q-4.0bpw	Instruct model fine-tuned using publicly available instruction / conversation datasets (no detailed breakdown publicly provided).
Llama3.1-8B-Instruct ^{17,22}	8B	General instruct	Open-weight, Q-4.0bpw	Pretrained on ~15T tokens from publicly available sources; fine-tuning includes public instruction datasets and >25M synthetic examples; cutoff December 2023.
gemma-2-9b-it ^{23,24}	9B	General instruct	Open-weight, Q-4.0bpw	Trained on 8T tokens (primarily English) from a mix including web documents, code, and science articles.
Llama-3-UltraMedical-8B ^{25,26}	8B	Biomedical-tuned	Open-weight, Q-4.0bpw Base: <i>Llama 3-8B</i>	Fine-tuned on UltraMedical: ~410k biomedical instruction samples + over 100k preference data (SFT + preference learning such as DPO/KTO).
Qwen3-30B-A3B ^{27,28}	30B total 3B active	General instruct	Open-weight, MoE, Q-4.0bpw	Qwen3 pretraining: 36T tokens across 119 languages, with a richer mix (coding, STEM, reasoning, books, multilingual, synthetic); Qwen3-30B-A3B released as “pretraining & post-training”.
Llama3.1-70B-Instruct ^{17,18}	70B	General instruct	Open-weight, Q-4.0bpw	Same Llama 3.1 recipe: ~15T public tokens + public instruction datasets + over 25M synthetic examples; cutoff Dec 2023.
Meditron3-70B ^{29,30}	70B	Clinical foundation (research-only; not instruction-tuned)	Open-weight, Q-4.0bpw Base: <i>Llama-3.1-70B</i>	Curated mixture: clinical guidelines, full-text medical articles, synthetic differential diagnosis conversations, replay general-domain data, and LLM-enhanced medical MCQs.

Eight locally deployable LLM backbones spanning 1B to 70B parameters, including general instruction-tuned, biomedical-tuned and clinical foundation models. Shown are model size, specialization, deployment notes and publicly described training-data composition. All models were evaluated under matched local 4-bit per weight quantized inference settings; full model-selection, deployment and inference details are provided in Methods, “Models and inference configuration”.

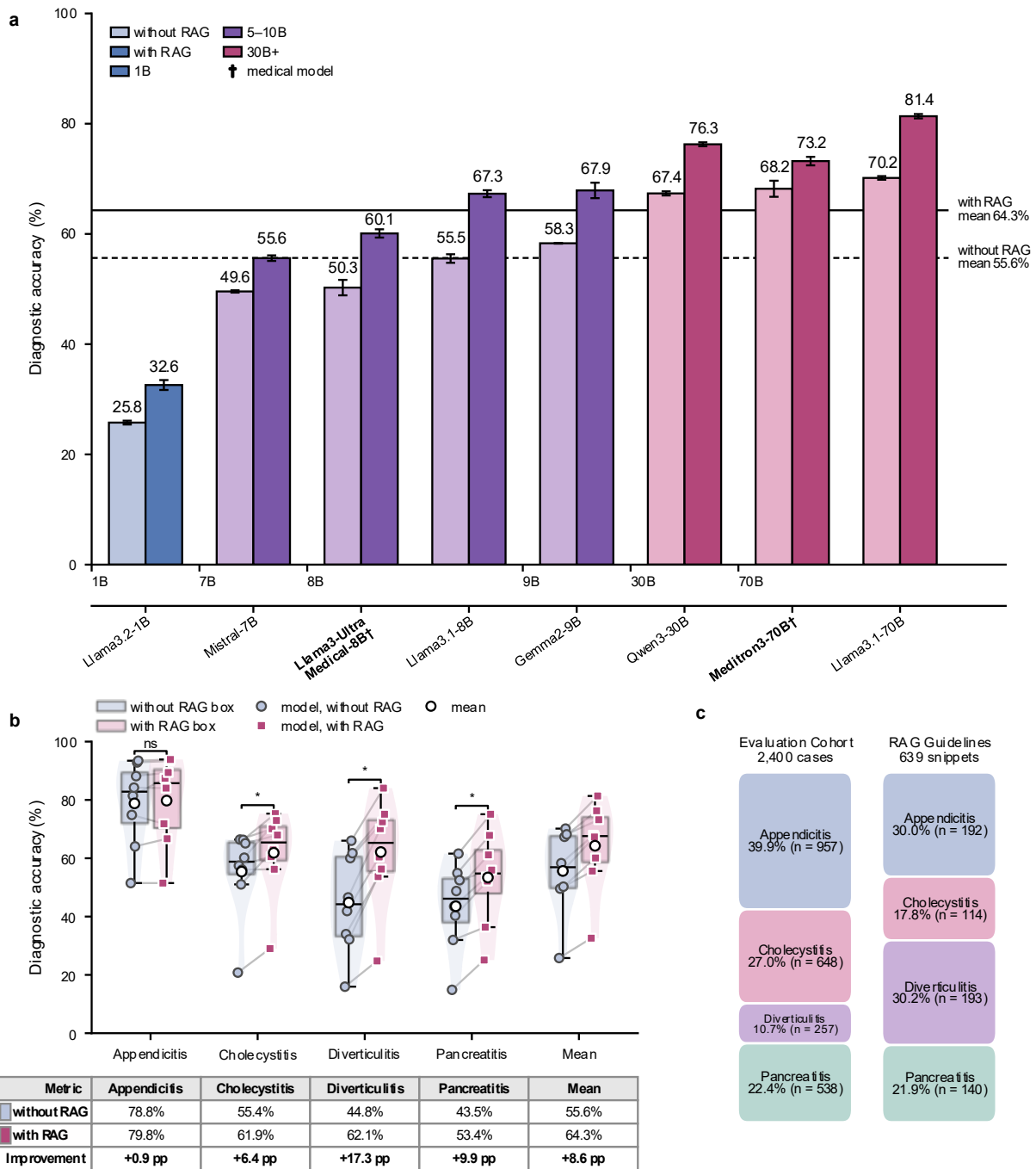


Fig. 3 | Retrieval-augmented generation improves diagnostic accuracy across model scale, domain and pathologies. **a**, Mean diagnostic accuracy for each of eight evaluated LLM backbones with and without RAG, macro-averaged across the four target pathologies. Paired bars show non-RAG and RAG accuracy, respectively; colors denote model-size groups and crosses denote medically adapted backbones. Values above bars indicate mean accuracy, and dashed horizontal lines show across-backbone means. Each backbone-condition was evaluated on all 2,400 MIMIC-CDM cases, across three decoding seeds. Error bars indicate sample standard deviations across the three seeds. **b**, Diagnostic accuracy stratified by pathology in matched non-RAG and RAG settings. Boxplots summarize model-level accuracies across the eight backbones; points denote individual backbones, averaged across three seeds, and labels indicate across-backbone means. Boxes span the interquartile range, center lines indicate medians and whiskers extend to the most extreme observations within $1.5 \times$ the interquartile range. The rightmost group shows macro-averaged accuracy across the four pathologies. The primary comparison in **a** used a one-sided Wilcoxon signed-rank test on paired model-level macro-averages ($n = 8$ backbones), under the prespecified alternative $RAG > \text{non-RAG}$; a one-sided paired t -test was used as a sensitivity analysis. Pathology-specific comparisons in **b** used one-sided Wilcoxon signed-rank tests on paired model-level accuracies ($n = 8$ backbones), with Holm adjustment across four comparisons. ns, adjusted $p \geq 0.05$; * adjusted $p < 0.05$. Further details are provided in Methods “Statistical analysis”.

We evaluated eight locally deployable LLM backbones ranging from 1B to 70B parameters, including general instruction-tuned, biomedical-adapted and clinical foundation models (Table 1). All models were run locally because MIMIC-CDM is derived from credentialed MIMIC-IV data that could not be transmitted to externally hosted model services. All models were run under matched 4-bit-per-weight quantized inference settings. Each backbone-condition combination was evaluated on all 2,400 cases across three decoding seeds.

Across the eight backbones, mean macro-averaged diagnostic accuracy increased from 55.64% (95% CI 43.53–67.76%) without retrieval to 64.29% (95% CI 51.52–77.07%) with retrieval. This corresponded to an absolute increase of 8.65 percentage points (95% CI 6.60–10.69; Fig. 3). For each backbone and condition, diagnostic accuracy was calculated over all 2,400 cases separately for each decoding seed and then averaged across the three seeds; these seed-averaged backbone estimates constituted the units for the model-level confidence intervals and paired hypothesis tests. The primary model-level comparison was significant using the prespecified one-sided Wilcoxon signed-rank test ($W = 36.00$, $P = 0.0039$). A one-sided paired t-test gave a concordant result ($t(7) = 10.00$, $P = 1.07 \times 10^{-5}$). These tests assess the consistency of paired improvements across the evaluated backbones. RAG increased mean accuracy across all eight backbones, with model-level absolute improvements ranging from 5.02 to 11.76 percentage points.

The improvement was distributed across model scale and training background rather than being driven by a single backbone (Fig. 3a). For example, accuracy increased from 25.77% to 32.59% for Llama3.2-1B^{17,19}, from 68.20% to 73.22% for the medically adapted Meditron3-70B^{29,30} and from 70.16% to 81.36% for Llama3.1-70B^{17,18}, corresponding to absolute improvements of 6.82, 5.02 and 11.20 percentage points, respectively. Complete model-level estimates are shown in Fig. 3a.

Importantly, the improvement was not restricted to smaller or lower-performing models. Among the three backbones containing at least 30B total parameters, mean accuracy increased from 68.57% (95% CI 64.99–72.15%) to 76.95% (95% CI 66.73–87.16%), an absolute increase of 8.38 percentage points (95% CI 0.62–16.13). Benefits were also observed across both general-purpose and medically adapted backbones: mean accuracy increased from 54.45% to 63.50% among the six general-purpose models and from 59.23% to 66.65% among the two medically adapted or clinical models. These findings indicate that external guideline evidence complemented, rather than merely compensated for, model scale and domain-specific pretraining.

When diagnostic accuracy was stratified by pathology, the magnitude of improvement varied across conditions (Fig. 3b). Across-backbone mean accuracy, calculated from model-level estimates averaged across the three decoding seeds, increased from 78.81% to 79.76% for appendicitis (difference, 0.95 percentage points; 95% CI –1.21 to 3.11; Holm-adjusted $P = 0.148$), from 55.42% to 61.87% for cholecystitis (difference, 6.45 percentage points; 95% CI 2.58–10.32; adjusted $P = 0.0156$), from 44.80% to 62.12% for diverticulitis (difference, 17.33 percentage points; 95% CI 12.81–21.85; adjusted $P = 0.0156$) and from 43.55% to 53.41% for pancreatitis (difference, 9.86 percentage points; 95% CI 6.64–13.08; adjusted $P = 0.0156$). Accordingly, the paired comparison was not significant for appendicitis after multiplicity adjustment, whereas significant improvements were observed for cholecystitis, diverticulitis and pancreatitis. The comparatively small change for appendicitis coincided with higher non-RAG performance, whereas the largest absolute improvement was observed for diverticulitis. Whether this heterogeneity reflects differences in baseline task difficulty, clinical presentation heterogeneity or the relevance of the retrieved guideline evidence should be interpreted cautiously.

Together, these results establish the primary finding of the study: guideline-grounded retrieval improved diagnostic accuracy across a heterogeneous range of clinical pathologies and locally deployable LLM backbones in a task requiring models to autonomously gather and interpret diagnostic evidence.

Clinical benefit depends on retrieval design and corpus composition

Table 2 | Overview of evidence corpora used for the corpus-ambiguity analysis.

Corpus	Item Count	Pathology						Notes
		Total	app.	chol.	div.	pan.	other	
Main four guidelines <i>Raw PDF</i>	Documents	4	1	1	1	1		The four guidelines used in the original MIMIC-CDM study to evaluate whether the LLMs fail to order the exams required by diagnostic guidelines. Guideline details are provided in Extended Data “Ranking Abdominal-Pain Diagnoses and Detailed Corpora Sources”.
	Snippets	422	158 (37.4%)	104 (24.6%)	85 (20.1%)	75 (17.8%)		
	Words	98,858	38,444	24,223	18,601	17,590		
Augmented corpus <i>Raw PDF</i>	Documents	16	4	4	4	4		Reusing the previous corpus as the base, adding 3 new guidelines from various sources per pathology. Guideline details are provided in Extended Data “Ranking Abdominal-Pain Diagnoses and Detailed Corpora Sources”.
	Snippets	980	310 (31.6%)	222 (22.7%)	249 (25.4%)	199 (20.3%)		
	Words	217,015	70,299	47,716	53,933	45,607		
Augmented corpus <i>Cleaned Markdown</i>	Documents	16	4	4	4	4		Baseline corpus for this study. Similar documents to previous corpus, but hand-edited for OCR noise and rewritten figures so the LLM can parse them. The cleaning process is described in Methods “Document cleaning, normalization and chunking”.
	Snippets	639	192 (30.0%)	114 (17.8%)	193 (30.2%)	140 (21.9%)		
	Words	153,463	52,020	33,716	35,912	31,815		
Ambiguous corpus <i>Raw PDF</i>	Documents	32	4	4	4	4	16	Reusing augmented corpus as the base, adding 16 guidelines covering other causes of abdominal pain, increasing diagnostic distractors in search and retrieval. Ambiguity selection and guideline details are described in Methods “Guideline corpus assembly”, and Extended Data “Ranking Abdominal-Pain Diagnoses and Detailed Corpora Sources”.
	Snippets	2817	310 (11.0%)	222 (7.9%)	249 (8.8%)	199 (7.1%)	1837 (65.2%)	
	Words	600,673	70,299	47,716	53,933	45,607	383,118	

Four retrieval corpora were evaluated, ranging from the original MIMIC-CDM guideline set to augmented, cleaned and ambiguity-enriched variants. Shown are source-document counts, indexed snippet counts, approximate total word count and pathology coverage for each corpus. The approximate total word count is computed as the total number of words contained within each snippet, including word overlap. The ambiguous corpus adds non-target abdominal-pain guidelines to stress test whether RAG remains beneficial when retrieval evidence becomes less pathology-specific. Full corpus construction, cleaning, ambiguity selection and chunking details are provided in Methods, “Guideline corpus assembly” and “Document cleaning, normalization and chunking”, and Extended Data “Ranking Abdominal-Pain Diagnoses and Detailed Corpora Sources”.

Baseline	Rank by median rank lower better ↓	Diagnostic accuracy % higher better ↑	Test-ordering alignment % higher better ↑	Treatment alignment % higher better ↑	Retrieval P@6 % higher better ↑	Retrieval spread (Brier score) lower better ↓
w/o RAG	10	59.58 ± 1.12	54.73 ± 0.51	42.03 ± 1.45	—	—
Corpora						
Ambiguous	9	65.91 ± 0.99	60.70 ± 0.80	42.10 ± 0.71	49.05 ± 0.86	0.375 ± 0.011
Main Four	7	66.26 ± 1.02	60.55 ± 0.40	43.39 ± 0.37	54.97 ± 0.60	0.309 ± 0.011
Augmented	3	68.58 ± 1.55	60.98 ± 0.12	43.68 ± 0.38	58.50 ± 0.14	0.268 ± 0.003
Ablations						
w/o requery	7	67.05 ± 0.93	59.33 ± 0.16	44.95 ± 0.39	47.96 ± 0.27	0.386 ± 0.005
w/o reranking	6	68.31 ± 1.55	59.22 ± 0.14	49.03 ± 0.30	55.59 ± 0.86	0.302 ± 0.011
w/o skip. 1st retrieval	5	69.63 ± 0.35	54.22 ± 0.98	42.31 ± 0.43	56.63 ± 0.71	0.297 ± 0.006
raw guidelines	3	68.58 ± 1.55	60.98 ± 0.12	43.68 ± 0.38	58.50 ± 0.14	0.268 ± 0.003
w/o u-shape ordering	1	70.21 ± 0.61	60.20 ± 0.50	46.37 ± 1.12	59.02 ± 0.82	0.266 ± 0.011
Our Framework						
All	1	69.78 ± 1.71	59.90 ± 1.25	46.00 ± 0.72	59.08 ± 0.84	0.266 ± 0.010

Fig. 4 | Clinical RAG depends on corpora and structured retrieval design. Performance of the full clinical RAG configuration (“All”), component ablations and corpora, with the matched non-RAG comparator. Results are averaged across *gemma-2-9b-it*^{23,24}, *Llama-3-UltraMedical-8B*^{25,26}, and *Llama3.1-70B-Instruct*^{17,18} with, each evaluated on all 2,400 cases across three seeds each. For each seed, metrics were averaged across backbones; means and 95% confidence intervals were then calculated across the three seed-level estimates ($n = 3$; two-sided Student’s *t* intervals). Therefore intervals quantify stochastic run-to-run variability of model-averaged performance. Table values are mean ± 95% CI half-width. Outcomes are diagnostic accuracy, test-ordering alignment, treatment alignment, retrieval precision at six chunks (P@6) and retrieval Brier score; lower Brier scores indicate retrieval distributions more concentrated on the reference pathology. “raw guidelines” removes OCR parsing and manual cleaning and instead uses the raw documents. “w/o u-shape ordering” removes U-shaped evidence ordering, which places the highest-ranked passages at the beginning and end of the evidence block to mitigate the lost-in-the-middle effect¹². “w/o skip. 1st retrieval” performs retrieval before the initial action from history alone; in “All”, retrieval begins after the first returned observation. “Augmented” and “raw guidelines” use the same 16-document raw corpus but are shown in the corpus-comparison and ablation sections, respectively, to make the controlled comparisons explicit. “All” uses the cleaned, structure-aware version of these same 16 source documents with the complete retrieval pipeline. Exact variant definitions and metric specifications are provided in Methods.

We next tested whether the observed benefit depended on how the guideline corpus and retrieval pipeline were constructed. Four corpora differed in scope, preparation and diagnostic ambiguity (Table 2). The main four-guideline corpus contained four pathology-specific documents and 422 indexed snippets. The augmented raw corpus included 16 documents and 980 snippets. Applying manual correction and structure-aware chunking to the same 16-document collection produced 639 snippets. Finally, the ambiguous corpus added 16 guidelines covering non-target causes of abdominal pain that were selected with the help of clinician expertise, resulting in 32 documents and 2,817 snippets. Because the raw and cleaned corpora used different conversion and segmentation procedures, their snippet counts should not be interpreted as direct measures of information volume.

The corpus series separated three properties of the evidence source: pathology coverage, document preparation and diagnostic ambiguity. The raw and cleaned augmented corpora contained the same source documents, allowing their comparison to isolate the effect of preprocessing. By contrast, the ambiguous corpus expanded the diagnostic search space with clinically plausible non-target causes of abdominal pain.

Test-ordering alignment measured the Bray-Curtis similarity between investigations requested by the agent and those performed by clinicians across physical examination, imaging and pathology-specific laboratory panels, with scores macro-averaged across pathologies. Treatment alignment quantified end-to-end concordance with clinician-recorded treatment decisions, defined as the proportion of clinician-derived expected treatment for which the agent both reached the correct diagnosis and requested the corresponding treatment, micro-averaged across patient-treatment pairs.

Retrieval outcomes were treated as intermediate characteristics of the evidence supplied to the agent rather than as clinical endpoints. Precision at six retrieved passages (P@6) measured the proportion, at every thinking step of the agent, of the six returned passages assigned to the reference pathology. The retrieval Brier score captured how retrieval mass was distributed across pathology labels, with lower values indicating greater concentration on the reference pathology.

Across the evaluated configurations, the complete retrieval framework showed one of the strongest overall performance profiles and tied for the best rank by median metric rank (rank 1; Fig. 4). The complete framework achieved a diagnostic accuracy of 69.78% (95% CI 68.07–71.49%), retrieval P@6 of 59.08% (95% CI 58.24–59.92%), a Brier score of 0.266 (95% CI 0.256–0.276), test-ordering alignment of 59.90% (95% CI 58.65–61.15%) and treatment alignment of 46.00% (95% CI 45.28–46.72%). Because the ablation analysis was designed to assess the repeatability of configuration-level performance for a fixed evaluation panel, metrics were averaged across backbones within each corresponding decoding seed and uncertainty was estimated across the three resulting seed-level means. The confidence intervals therefore quantify stochastic run-to-run variability of model-averaged performance rather than estimates of inter backbone heterogeneity. The limited seed-level variation indicates that the aggregated configuration estimates were reproducible across the three decoding runs. Although the complete framework produced a consistently favorable overall profile, it was not the highest-performing configuration for every individual metric.

Query rewriting was particularly consequential for retrieval fidelity. Removing it meant that the complete patient context was used directly as the search query. Compared with the complete framework, removing query rewriting reduced diagnostic accuracy from 69.78% to 67.05%, a decrease of 2.73 percentage points; test-ordering alignment from 59.90% to 59.33%, a decrease of 0.57 percentage points; treatment alignment from 46.00% to 44.95%, a decrease of 1.05 percentage points; and P@6 from 59.08% to 47.96%, a decrease of 11.12 percentage points. The Brier score increased from 0.266 to 0.386, a difference of 0.120. Among the individual component ablations, removing query rewriting therefore produced the largest deterioration in both retrieval P@6 and retrieval spread while also reducing each of the reported clinical performance measures. These findings indicate that the accumulated patient record was not an effective retrieval query by itself and that autonomously compressing it into a focused representation of the current clinical problem improved pathology-relevant retrieval.

Reranking provided a further contribution to retrieval performance. Removing reranking reduced P@6 from 59.08% to 55.59% and increased the Brier score from 0.266 to 0.302. These retrieval changes were accompanied by a reduction in diagnostic accuracy from 69.78% to 68.31%, a difference of 1.47 percentage points, and in test-ordering alignment from 59.90% to 59.22%, a difference of 0.68 percentage points. In contrast, treatment alignment increased from 46.00% to 49.03% when reranking was removed. Thus, reranking improved retrieval fidelity and was associated with higher diagnostic accuracy and test-ordering alignment, although its effect was not uniformly favorable across all downstream clinical outcomes.

The corpus comparisons showed that broader target-pathology coverage was beneficial when the added material remained sufficiently specific and retrievable. The compact four-guideline corpus underperformed the raw (not manually cleaned) augmented corpus across all reported outcomes: diagnostic accuracy was 66.26% versus 68.58%, test-ordering alignment was 60.55% versus 60.98%, treatment alignment was 43.39% versus 43.68%, P@6 was 54.97% versus 58.50%, and the Brier score was 0.309 versus 0.268. Thus, expanding the corpus from one to four guidelines per target pathology improved retrieval fidelity and was accompanied by modest improvements in downstream clinical performance, although the value of further expansion depended on the diagnostic composition of the added evidence.

The ambiguous corpus was deliberately constructed as a retrieval stress test by adding 16 guidelines for non-target abdominal diagnoses with overlapping symptoms and diagnostic pathways. Relative to the raw augmented corpus, diagnostic accuracy decreased from 68.58% to

65.91%, P@6 from 58.50% to 49.05%, test-ordering alignment from 60.98% to 60.70% and treatment alignment from 43.68% to 42.10%, while the Brier score increased from 0.268 to 0.375. Nevertheless, ambiguous-corpus RAG remained better than no RAG for the principal diagnostic and test-ordering outcomes, increasing diagnostic accuracy from 59.58% to 65.91%, an absolute gain of 6.33 percentage points, and test-ordering alignment from 54.73% to 60.70%, a gain of 5.97 percentage points. Moreover, although each target pathology represented only 7.1–11.0% of the indexed corpus, 49.05% of retrieved passages corresponded to the reference pathology. Thus, deliberately introduced diagnostic ambiguity impaired retrieval fidelity but neither overwhelmed the retrieval pipeline nor eliminated its clinical benefit. Pathology-resolved retrieval distributions for the clean augmented and ambiguous-corpus configurations are shown in Extended Data Fig. 2 “Composition of guideline retrieval across patient pathologies and source documents from the baseline augmented corpus with cleaned markdown” and Extended Data Fig. 3 “Composition of guideline retrieval across patient pathologies and source documents from the ambiguous corpus”.

Allowing retrieval before the first returned observation had its largest effect on test-ordering alignment, which decreased from 59.90% to 54.22%, a difference of 5.68 percentage points. Treatment alignment decreased from 46.00% to 42.31%, a difference of 3.69 percentage points; diagnostic accuracy decreased by 0.15 percentage points (69.78% to 69.63%); P@6 decreased by 2.45 percentage points (59.08% to 56.63%); and the Brier score increased from 0.266 to 0.297, a difference of 0.031. This identifies retrieval timing as an additional design variable. The initial history may not contain enough specific information to select reliable disease-focused guidance. Although this does not establish an anchoring mechanism, it is consistent with the possibility that prematurely retrieving information without enough clinical information, as before the first step, not even a physical examination of the patient has been conducted yet, can lead to baseless retrieved passages, influencing the agent requests next and, consequently, which clinical information becomes available later.

Retrieval fidelity and downstream clinical performance were related, but not interchangeable. Configurations retrieving more pathology-concentrated evidence generally achieved higher diagnostic accuracy and closer alignment with clinician test ordering, but the configuration with the strongest retrieval scores was not necessarily the best for every process-level outcome. Overall, these experiments show that the clinical benefit of RAG arose from a coordinated retrieval system, particularly autonomous task-focused query rewriting and passage-level reranking, rather than from simply appending more retrieved text to the model context.

Discussion

In this study, guideline-grounded retrieval improved diagnostic accuracy across eight LLM backbones during iterative clinical decision-making, increasing mean macro-averaged accuracy from 55.64% without retrieval to 64.29% with retrieval, an absolute improvement of 8.65 percentage points (95% CI 6.60–10.69). In contrast to single-turn evaluations, MIMIC-CDM required agents to select examinations, integrate returned findings and autonomously decide when sufficient evidence supported a diagnosis, so retrieval influenced an evolving sequence of decisions rather than only a final response. Improvements were observed across all eight backbones but varied by pathology, with the largest gains for diverticulitis and pancreatitis and a marginal improvement for appendicitis. These findings indicate that inference-time guideline evidence can improve diagnostic performance despite encoded medical knowledge, advancing agent transparency, although they do not establish clinical safety or effectiveness.

The improvement in final diagnosis was greater than that in the clinical pathway leading to it. In the matched three-backbone analysis, diagnostic accuracy increased by 10.20 percentage points relative to non-RAG, compared with 5.17 points for test-ordering alignment and 3.97 points for treatment alignment. A model may therefore reach the correct diagnosis despite an incomplete or potentially inappropriate work-up; endpoint accuracy should not be interpreted as evidence that preceding actions were safe or optimal. Our findings extend rather than overturn the original MIMIC-CDM study, which identified limitations in instruction following and workflow behavior^{15,31}. Guideline grounding mitigated part of this failure profile while increasing traceability, but did not resolve workflow and safety barriers.

The central mechanistic finding was the importance of autonomous query rewriting. As each case progressed, the context accumulated history, examination findings, reports and previous model outputs that were not equally informative for retrieval. Replacing the task-focused rewritten query with accumulated context reduced P@6 from 59.08% to 47.96% and increased the retrieval Brier score from 0.266 to 0.386. Effective retrieval therefore appeared to require an explicit representation of the agent’s current decision state rather than embedding the entire evolving record. Query rewriting can be viewed as a compression step that converts heterogeneous and verbose context into the current clinical information need. Because it may also omit relevant information or prematurely narrow the differential, rewritten queries should remain auditable.

Reranking and retrieval timing provided additional evidence that retrieval quality depends on how and when external knowledge is introduced. Candidate retrieval determines which passages can enter consideration, reranking which receive the limited evidence budget, and timing how much clinical information is available when searching. Removing reranking worsened retrieval fidelity and modestly reduced diagnostic accuracy and test-ordering alignment, although effects were not uniform across outcomes. Allowing retrieval before the first returned observation reduced test-ordering alignment by 5.68 percentage points while leaving diagnostic accuracy nearly unchanged. Evidence prioritization and retrieval scheduling should therefore be treated as design variables, although these findings do not establish a specific causal mechanism.

Corpus composition showed that sensitivity to the retrieval environment and robustness to retrieval noise are not contradictory. Expanding guideline coverage across target pathologies improved retrieval fidelity and diagnostic performance, whereas adding plausible, symptomatically overlapping non-target stress-test guidelines reduced both. Nevertheless, ambiguous-corpus RAG remained above matched non-RAG for diagnostic accuracy and test-ordering alignment, and nearly half of retrieved passages still corresponded to the reference pathology despite each target pathology representing only 7.1–11.0% of the index. Corpus ambiguity therefore attenuated rather than eliminated the observed benefit. Corpus curation is therefore part of the design variables, supporting corpus governance, version control and jurisdictional metadata.

Together, these experiments show that retrieval fidelity and downstream clinical performance are related but not interchangeable. Improvements in retrieval metrics did not translate monotonically to every process-level outcome, and diagnostic accuracy alone would have obscured changes in both the evidence supplied to the model and the pathway to the final diagnosis. “RAG” is consequently not a sufficiently precise design description unless corpus construction, query generation, reranking, evidence organization and retrieval scheduling are reported. Retrieved sources, rewritten queries and intermediate actions should likewise remain available for audit^{7,13,32,33}.

Several limitations should be considered. MIMIC-CDM derives from a single US academic medical center, includes only four abdominal pathologies and uses retrospective records and a constrained action space. It does not reproduce clinical time pressure, local resource constraints, clinician-patient communication or consequences of delayed or unnecessary testing. Test-ordering alignment and treatment alignment reflect similarity to investigations recorded in source data rather than proof of an optimal work-up, while pathology-labelled retrieval relevance does not capture recommendation applicability, evidence strength or guideline conflicts. Generalizability to other institutions, languages, guideline jurisdictions, multimodal systems and future model versions remains uncertain. Statistical uncertainty differed by analysis. In the primary eight-backbone analysis, seeds were averaged within backbones and CIs calculated across backbones, reflecting model heterogeneity. In the retrieval-design experiments, the fixed three-backbone panel was averaged within each seed and CIs calculated across seeds, reflecting stochastic run-to-run variability rather than generalizability. Finally, the study does not establish treatment safety, triage performance, effects on clinician decisions, patient outcomes or automation bias.

The next step is external, prospective evaluation under clinician oversight. Silent-mode studies across institutions and guideline jurisdictions could assess diagnoses, retrieved evidence, omitted work-up and potentially unsafe recommendations. Reader studies could then determine whether evidence-linked recommendations improve clinician performance without prematurely narrowing the differential or increasing automation bias. Subsequent pragmatic or randomized studies would be required to establish effects on workflow and patient-centered outcomes. Prospective evaluation should also monitor corpus updates, conflicting recommendations, query-rewriting failures, subgroup performance and model-version drift.

In conclusion, guideline-grounded retrieval improved diagnostic performance in a real-patient benchmark of iterative clinical decision-making, but its benefit depended on retrieval design. Autonomous query rewriting produced the clearest ablation signal, particularly for retrieval fidelity, while reranking, corpus composition and retrieval timing altered the evidence supplied to the agent and selected downstream outcomes. These findings support structured retrieval, rather than retrieval alone, as the relevant object of evaluation for guideline-grounded clinical agents. They support further evaluation of RAG as supervised, auditable clinical decision support, but do not support autonomous clinical deployment.

Online Content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <DOI>.

References

1. Wang, D. & Zhang, S. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial Intelligence Review* **57**, (2024).
2. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nature Medicine* **29**, 1930–1940 (2023).
3. Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and adoption of large language models in medicine. *JAMA* **330**, 866–869 (2023).
4. Huang, L. *et al.* A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. (2023).
5. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* 610–623 (Association for Computing Machinery, 2021). doi:10.1145/3442188.3445922.
6. Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. in *Advances in neural information processing systems* vol. 33 9459–9474 (2020).
7. Zakka, C. *et al.* Almanac — Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI* **1**, Aloa2300068 (2024).
8. Gao, Y. *et al.* Large Language Models and Medical Knowledge Grounding for Diagnosis Prediction. Preprint at <https://doi.org/10.1101/2023.11.24.23298641> (2023).
9. Yu, H. *et al.* Evaluation of retrieval-augmented generation: a survey. (2024).
10. Wyer, P. C., Gabbay, J. & Suh, E. H. Why all clinical guideline recommendations are ‘Conditional’. *Clinical and Public Health Guidelines* **1**, e12013 (2024).
11. Yao, L. *et al.* Discordant and inappropriate discordant recommendations in consensus and evidence based guidelines: empirical analysis. *BMJ* **375**, e066045 (2021).
12. Liu, N. F. *et al.* Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* **12**, 157–173 (2024).

13. Xiong, G., Jin, Q., Lu, Z. & Zhang, A. Benchmarking Retrieval-Augmented Generation for Medicine. in *Findings of the Association for Computational Linguistics: ACL 2024* (eds Ku, L.-W., Martins, A. & Srikumar, V.) 6233–6251 (Association for Computational Linguistics, Bangkok, Thailand, 2024). doi:10.18653/v1/2024.findings-acl.372.
14. Feldman, F. *et al.* Healthier LLMs: Retrieval-Augmented Generation for Public Health Question Answering. Preprint at <https://doi.org/10.48550/arXiv.2607.06641> (2026).
15. Hager, P. *et al.* Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine* **30**, 2613–2622 (2024).
16. Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* **10**, 1 (2023).
17. Grattafiori, A. *et al.* The llama 3 herd of models. (2024) doi:10.48550/arXiv.2407.21783.
18. Meta. Llama-3.1-70B-instruct model card. (2024).
19. Meta. Llama-3.2-1B-instruct model card. (2024).
20. Jiang, A. Q. *et al.* Mistral 7B. (2023) doi:10.48550/arXiv.2310.06825.
21. Mistral AI. Mistral-7B-instruct-v0.3 model card. (2024).
22. Meta. Llama-3.1-8B-instruct model card. (2024).
23. Gemma Team *et al.* Gemma 2: Improving open language models at a practical size. (2024) doi:10.48550/arXiv.2408.00118.
24. Google. gemma-2-9b-it model card. (2024).
25. Zhang, K. *et al.* UltraMedical: Building specialized generalists in biomedicine. in *Advances in neural information processing systems* vol. 37 26045–26081 (2024).
26. Tsinghua C3I Lab. Llama-3-8B-UltraMedical model card. (2024).
27. Qwen Team. Qwen3 technical report. (2025) doi:10.48550/arXiv.2505.09388.
28. Qwen. Qwen3-30B-A3B model card. (2025).
29. Sallinen, A. *et al.* Llama-3-meditron: An open-weight suite of medical llms based on llama-3.1. in *AAAI 2025 workshop on large language models and generative AI for health* (2025).

30. OpenMeditron Initiative. Meditron3-70B model card. (2025).
31. Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data* **10**, 1 (2023).
32. Gallifant, J. *et al.* The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine* <https://doi.org/10.1038/s41591-024-03425-5> (2025) doi:10.1038/s41591-024-03425-5.
33. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine* **26**, 1364–1374 (2020).
34. PhysioNet. MIMIC-IV. (2024).
35. PhysioNet. PhysioNet credentialed health data license and data use agreement. (2024).
36. Auer, C. *et al.* Docling technical report. (2024).
37. Robertson, S. & Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* **3**, 333–389 (2009).
38. NovaSearch Team. Jasper and stella: Distillation of SOTA embedding models. (2024).
39. Zhang, Y. *et al.* Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. Preprint at <https://doi.org/10.48550/arXiv.2506.05176> (2025).
40. Lee, C. *et al.* NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. Preprint at <https://doi.org/10.48550/arXiv.2405.17428> (2025).
41. Jin, Q. *et al.* MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics* **39**, btad651 (2023).
42. Douze, M. *et al.* The faiss library. (2024).
43. Sentence Transformers. cross-encoder/ms-marco-MiniLM-L6-v2 model card. (2026).
44. Gallego, F., Ruas, P., Couto, F. M. & Veredas, F. J. Enhancing cross-encoders using knowledge graph hierarchy for medical entity linking in zero- and few-shot scenarios. *Knowledge-Based Systems* **314**, 113211 (2025).

45. BSC-NLP4BIA/Medprocner-CE-Reranker · Hugging Face. <https://huggingface.co/BSC-NLP4BIA/Medprocner-CE-Reranker> (2025).
46. Yoran, O., Wolfson, T., Ram, O. & Berant, J. Making retrieval-augmented language models robust to irrelevant context. (2023).
47. Asai, A., Wu, Z., Wang, Y., Sil, A. & Hajishirzi, H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. in *International conference on learning representations* (2024).
48. Jiang, Z. *et al.* Active retrieval augmented generation. in *Proceedings of the 2023 conference on empirical methods in natural language processing* 7969–7992 (2023). doi:10.18653/v1/2023.emnlp-main.495.
49. Mamede, S. *et al.* Role of knowledge and reasoning processes as predictors of resident physicians' susceptibility to anchoring bias in diagnostic reasoning: a randomised controlled experiment. *BMJ Quality & Safety* **33**, 563–572 (2024).
50. turboderp-org. ExLlamaV2: A fast inference library for running LLMs locally on modern consumer-class GPUs. (2024).
51. Qwen. Qwen3-embedding-0.6B model card. (2025).
52. Snyder, M. J., Guthrie, M. & Cagle, S. Acute Appendicitis: Efficient Diagnosis and Management. *AFP* **98**, 25–33 (2018).
53. Di Saverio, S. *et al.* Diagnosis and treatment of acute appendicitis: 2020 update of the WSES Jerusalem guidelines. *World J Emerg Surg* **15**, 27 (2020).
54. Kumar, S. S. *et al.* SAGES guideline for the diagnosis and treatment of appendicitis. *Surg Endosc* **38**, 2974–2994 (2024).
55. Bonomo, R. A. *et al.* 2024 Clinical Practice Guideline Update by the Infectious Diseases Society of America on Complicated Intra-abdominal Infections: Diagnostic Imaging of Suspected Acute Appendicitis in Adults, Children, and Pregnant People. *Clin Infect Dis* **79**, S94–S103 (2024).

56. *Diverticular Disease: Diagnosis and Management*. (National Institute for Health and Care Excellence (NICE), London, 2019).
57. Hall, J. *et al.* The American Society of Colon and Rectal Surgeons Clinical Practice Guidelines for the Treatment of Left-Sided Colonic Diverticulitis. *Dis Colon Rectum* **63**, 728–747 (2020).
58. Peery, A. F., Shaukat, A. & Strate, L. L. AGA Clinical Practice Update on Medical Management of Colonic Diverticulitis: Expert Review. *Gastroenterology* **160**, 906-911.e1 (2021).
59. Schultz, J. K. *et al.* European Society of Coloproctology: guidelines for the management of diverticular disease of the colon. *Colorectal Dis* **22 Suppl 2**, 5–28 (2020).
60. Goodchild, G., Chouhan, M. & Johnson, G. J. Practical guide to the management of acute pancreatitis. *Frontline Gastroenterol* **10**, 292–299 (2019).
61. National Guideline Centre (Great Britain). *Pancreatitis: Final Guideline*. (National Institute for Health and Care Excellence, London, 2018).
62. Leppäniemi, A. *et al.* 2019 WSES guidelines for the management of severe acute pancreatitis. *World J Emerg Surg* **14**, 27 (2019).
63. Working Group IAP/APA Acute Pancreatitis Guidelines. IAP/APA evidence-based guidelines for the management of acute pancreatitis. *Pancreatology* **13**, e1-15 (2013).
64. ten Broek, R. P. G. *et al.* Bologna guidelines for diagnosis and management of adhesive small bowel obstruction (ASBO): 2017 update of the evidence-based guidelines from the world society of emergency surgery ASBO working group. *World J Emerg Surg* **13**, 24 (2018).
65. Diagnosis and Management of Ectopic Pregnancy. *BJOG: An International Journal of Obstetrics & Gynaecology* **123**, e15–e55 (2016).
66. Bala, M. *et al.* Acute mesenteric ischemia: updated guidelines of the World Society of Emergency Surgery. *World J Emerg Surg* **17**, 54 (2022).

67. Wanhainen, A. *et al.* Editor's Choice -- European Society for Vascular Surgery (ESVS) 2024 Clinical Practice Guidelines on the Management of Abdominal Aorto-Iliac Artery Aneurysms. *Eur J Vasc Endovasc Surg* **67**, 192–331 (2024).
68. Mencarini, L., Vestito, A., Zagari, R. M. & Montagnani, M. The Diagnosis and Treatment of Acute Cholecystitis: A Comprehensive Narrative Review for a Practical Approach. *J Clin Med* **13**, 2695 (2024).
69. Pisano, M. *et al.* 2020 World Society of Emergency Surgery updated guidelines for the diagnosis and treatment of acute calculus cholecystitis. *World J Emerg Surg* **15**, 61 (2020).
70. G, B. *et al.* Gallstone Disease - NICE Quality Standard [QS104]. (2015). doi:10.13140/RG.2.2.20305.53602.
71. Yokoe, M. *et al.* Tokyo Guidelines 2018: diagnostic criteria and severity grading of acute cholecystitis (with videos). *J Hepatobiliary Pancreat Sci* **25**, 41–54 (2018).
72. Skolarikos, Andreas *et al.* EAU Guidelines on Urolithiasis. <https://uroweb.org/guidelines/urolithiasis/chapter/guidelines> (2025).
73. Buxbaum, J. L. *et al.* ASGE guideline on the management of cholangitis. *Gastrointest Endosc* **94**, 207-221.e14 (2021).
74. Lebedeva, M. *Protocol for Diagnosis and Treatment of Peptic Ulcer in Adults*. <https://www.aiha.com/wp-content/uploads/2015/07/CPGUlcerEN.pdf> (2002).
75. Bishop, E. J. & Tiruvoipati, R. Management of Clostridioides difficile infection in adults and challenges in clinical practice: review and comparison of current IDSA/SHEA, ESCMID and ASID guidelines. *J Antimicrob Chemother* **78**, 21–30 (2023).
76. Singh, S. *et al.* AGA Living Clinical Practice Guideline on Pharmacological Management of Moderate-to-Severe Ulcerative Colitis. *Gastroenterology* **167**, 1307–1343 (2024).
77. *Pyelonephritis (Acute): Antimicrobial Prescribing*. (National Institute for Health and Care Excellence (NICE), London, 2018).

78. Dhatariya, K. K. & Care, T. J. B. D. S. for I. The management of diabetic ketoacidosis in adults—An updated guideline from the Joint British Diabetes Society for Inpatient Care. *Diabetic Medicine* **39**, e14788 (2022).
79. Kim, Y. J. *et al.* Guideline for the Antibiotic Use in Acute Gastroenteritis. *Infection & Chemotherapy* **51**, 217–243 (2019).
80. Sharp, V. J., Kieran, K. & Arlen, A. M. Testicular Torsion: Diagnosis, Evaluation, and Management. *AFP* **88**, 835–840 (2013).
81. Yusuf, H. & Trent, M. Management of Pelvic Inflammatory Disease in Clinical Practice. *Ther Clin Risk Manag* **19**, 183–192 (2023).
82. Hung, A., Calderbank, T., Samaan, M. A., Plumb, A. A. & Webster, G. Ischaemic colitis: practical challenges and evidence-based recommendations for management. *Frontline Gastroenterol* **12**, 44–52 (2021).
83. European Association for the Study of the Liver. EASL clinical practice guidelines on the management of ascites, spontaneous bacterial peritonitis, and hepatorenal syndrome in cirrhosis. *J Hepatol* **53**, 397–417 (2010).

Methods

Study design and reporting framework

We performed a retrospective benchmark study to quantify the effect of retrieval-augmented generation (RAG) on large language model (LLM) agents performing iterative clinical decision-making on the MIMIC Clinical Decision-Making benchmark (MIMIC-CDM)¹⁵. The primary endpoint was diagnostic accuracy. Secondary clinical outcomes were test-ordering alignment and treatment alignment. Retrieval performance was assessed using precision at six retrieved passages (P@6) and the retrieval Brier score, repeated for each thinking step on a case. The unit of analysis was the individual patient trajectory. Unless otherwise stated, each RAG configuration was compared to its matched non-RAG comparator using the same cases, backbone model, decoding settings and seeds. This study was not preregistered. Analysis of the primary endpoint was planned before full evaluation. Additional subgroup, ablation and sensitivity analyses were exploratory.

Source benchmark and case population

We used MIMIC-CDM¹⁵, a curated benchmark of 2,400 emergency-department presentations derived from the de-identified MIMIC-IV electronic health record corpus^{16,34}. MIMIC-IV contains structured and unstructured clinical data, including diagnoses, procedures, laboratory and microbiology measurements, medications, radiology reports and free-text notes, for patients treated at Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA¹⁶. In keeping with the original MIMIC-CDM benchmark, the present study focused on acute abdominal pain presentations whose final primary diagnosis was appendicitis, cholecystitis, diverticulitis or pancreatitis¹⁵. The benchmark comprises 957 appendicitis cases, 648 cholecystitis cases, 257 diverticulitis cases and 538 pancreatitis cases.

Cohort construction, case curation and reference diagnosis labeling were inherited from the original MIMIC-CDM release and were not redefined in the present work¹⁵. Briefly, the source benchmark filtered cases using diagnosis codes and manual review of discharge summaries to ensure that one of the four target abdominal pathologies was the first primary diagnosis and reason for emergency presentation, rather than an additional diagnosis during a extensive admission¹⁵. We retained the original inclusion and exclusion rules so that our results would remain directly comparable with the benchmark on which the present study builds.

Each case exposed the same information channels as in the original framework: history of present illness, physical examination, laboratory and microbiology results, radiology reports, and procedures performed during the admission¹⁵. Across the full benchmark, this corresponds to physical examinations for all 2,400 patients, 138,788 laboratory results from 480 unique laboratory tests, 4,403 microbiology results from 74 unique tests, and 5,959 radiology reports. Radiology reports include 1,836 abdominal CT examinations, 1,728 chest X-rays, 1,325 abdominal ultrasounds, 342 abdominal X-rays and 227 MRCP examinations.

We reused the original test-name normalization resources supplied with MIMIC-CDM, including laboratory synonym dictionaries and mappings from imaging exam names to modality and anatomical region¹⁵. We did not modify patient trajectories, labels or tool-return logic except where explicitly described below for the addition of retrieval.

Inherited evaluation framework and modifications in the present study

The interactive decision-making task was inherited from the original MIMIC-CDM framework¹⁵. In that framework, an LLM receives the history of present illness, iteratively requests further clinical information, and eventually outputs a final diagnosis and management plan. We retained the benchmark task definition, available tool types, case order within each experiment, and evaluation against the benchmark reference labels.

The methodological contribution of the present study was the addition of a RAG layer on top of this inherited benchmark. New components introduced here were: (1) assembly of guideline corpora, (2) document cleaning and normalization for machine retrieval, (3) lexical, dense and hybrid retrieval configurations, (4) optional cross-encoder reranking, (5) optional autonomous LLM-based query rewriting, (6) per-step evidence injection into the agent context, and (7) retrieval-focused and process-focused analyses not present in the original MIMIC-CDM paper. Accordingly, the benchmark and task were inherited, whereas the non-parametric evidence source and retrieval system were new.

Ethics, privacy and data governance

All analyses were performed inside an access-controlled computing environment. No patient-level data were transmitted to external application programming interfaces or third-party hosted model endpoints. This restriction followed the credentialed-access and data-governance requirements of MIMIC-IV and PhysioNet^{16,34,35}. For that reason, all evaluated backbone models were open-weight models deployed locally.

MIMIC-IV is a de-identified research resource released under credentialed access, and its source data collection and sharing were previously reviewed and approved with waiver of informed consent by the Beth Israel Deaconess Medical Center Institutional Review Board¹⁶. The present work used only de-identified benchmark data derived from that resource.

Guideline corpus assembly

The retrieval corpus was designed as the external, non-parametric evidence source used to ground model decisions. We manually curated adult clinical guidance documents relevant to the four benchmark pathologies and, in selected experiments, to additional abdominal-pain differentials. For each included document, we recorded the source organization, title, publication year or version, and pathology label in Extended Data “Ranking Abdominal-Pain Diagnoses and Detailed Corpora Sources”.

We evaluated multiple corpus variants. The base corpus comprised 4 source guidelines covering the four benchmark diseases, and corresponding to the four guidelines used in the original MIMIC-CDM study to evaluate whether the LLMs fail to order the exams required by diagnostic guidelines¹⁵. An expanded corpus comprised 16 documents, including additional guidance per target pathology, with 4 guidelines per target pathology. A higher-ambiguity corpus further added 16 documents (for a total of 32) covering 16 non-target causes of acute abdominal pain selected with the help of a clinician expert, introducing retrieval distractors, as a retrieval stress-test and better approximating differential diagnosis in the emergency department. Exact document lists, references, as well as the diagnostic-distractor selection logic is described in Extended Data “Ranking Abdominal-Pain Diagnoses and Detailed Corpora Sources”.

Document cleaning, normalization and chunking

Source guideline documents were converted into machine-readable text using Docling OCR plus manual correction³⁶. After conversion, documents underwent manual quality control to remove OCR artifacts, normalize whitespace and Unicode characters, and preserve only clinically meaningful section structure, removing sections such as authors, references, fundings, etc. Figures, flowcharts and non-machine-readable algorithms were manually rewritten into textual pseudo-code to ensure that diagnostically relevant content remained retrievable.

When possible, we preserved document hierarchy during chunking by retaining section headings and storing them as chunk-level metadata. Each chunk carried provenance fields including document source, document format, chunk ID, page number, order in document, and,

when available, heading-path metadata. Chunking depended on corpus format. For raw PDF and other non-Docling sources, documents were split into fixed-size overlapping passages of 512 tokens with 52-token overlap. For hand-cleaned guideline corpora converted to Markdown, we instead used Docling’s HybridChunker³⁶ with the embedding-model tokenizer as the token budget, so 512 tokens acted as an upper bound rather than a fixed chunk length. In this setting, Docling applied tokenization-aware refinements on top of document-based hierarchical chunking: splitting only chunks that exceed the token limit and merging undersized adjacent chunks when they shared compatible headings or captions, while preserving document structure and table context. Importantly, for these Markdown documents, we embedded the chunker’s contextualized text representation rather than the raw chunk text. Exact corpus composition, document counts, chunk counts, words counts, and tags for each corpus variant are provided in Table 2 and Extended Data “Ranking Abdominal-Pain Diagnoses and Detailed Corpora Sources”.

Retrieval stack

We evaluated selected lexical, dense and hybrid retrieval configurations in a dedicated retrieval-stack selection stage. Lexical retrieval used BM25³⁷. Candidate dense retrievers comprised stella_en_400M_v5³⁸, Qwen3-Embedding-0.6B³⁹, NV-Embed-v2⁴⁰ and MedCPT-Article/Query-Encoder⁴¹. Embeddings were precomputed for all chunks, cast to float32, and indexed with FAISS IndexFlatL2⁴². Dense candidates were retrieved by nearest-neighbor search under L2 distance in embedding space.

In hybrid configurations, BM25 (lexical) retrieval and dense (semantic) retrieval were run in parallel and their candidate lists were merged before optional reranking. Before merging, dense and BM25 scores were independently min-max normalized within each query and combined by weighted linear interpolation with $\alpha = 0.65$ (65% dense and 35% lexical score). Duplicate chunks were removed by unique chunk identifier. Candidate rerankers comprised cross-encoder/ms-marco-MiniLM-L6-v2⁴³, Qwen3-Reranker-0.6B³⁹, MedCPT-Cross-Encoder⁴¹ and Medprocner-CE-Reranker^{44,45}; configurations without a reranker were also evaluated. The default candidate pool before reranking comprised 60 lexical and 60 dense candidates, and the final evidence set returned to the agent contained the top-6 chunks.

Retrieval-stack selection was evaluated using Gemma2-9B and Llama3.1-70B, with each candidate configuration run on all 2,400 cases using one decoding seed. After completion of these retrieval-stack selection experiments, BM25 + Qwen3-Embedding-0.6B + MedCPT-Cross-Encoder was selected as the fixed default retrieval stack and used for all subsequent main experiments in the study, unless otherwise specified. Retrieval-stack selection and performance comparisons are reported in Extended Data Fig. 1, “Retrieval stack selection/performance comparison”.

Interactive RAG-enhanced benchmark execution

At each benchmark step, the agent operated on the currently accumulated patient state, consisting of the history of present illness and all clinical information (tests reports) returned up to that point. In the non-RAG comparator, this patient state was passed directly to the backbone LLM together with the task instructions and output schema. In the RAG condition, at each step, the current patient state first triggered a retrieval cycle, after which the retrieved evidence was used to augment the prompt before the model generated its next action or final diagnosis.

The available actions matched the inherited MIMIC-CDM task: physical examination, laboratory testing, imaging, or final diagnosis with management plan¹⁵. After each model output, the framework parsed the requested action, executed the corresponding tool if valid, returned the resulting observation, and advanced the case to the next step. This loop continued until the model finalized its answer or until a stopping rule described below was reached.

Query rewriting and evidence assembly

We evaluated two strategies for forming retrieval queries from the evolving patient context. In the naive condition, the retrieval input was the serialized case context itself, truncated if necessary to the input limit of the retrieval model. In the re-query condition, reusing the already loaded backbone LLM as the acting agent, via a separate parallel query-writing prompt, transformed the current patient context into a shorter, task-focused search query intended to preserve diagnostically relevant content while reducing verbosity; this auxiliary prompt had access only to the current patient state at that step and not to reference labels or future case information. The exact prompt template used for autonomous query rewriting is provided in Extended Data “Requery System Prompt”.

Retrieved passages were inserted into the model prompt as an evidence block following the task instructions and preceding the evolving patient context. Each retrieved passage was presented together with lightweight provenance information visible to the model, including simplified guideline title, year of issue, and complete headings hierarchy leading to the passage’s section. Full provenance metadata were also stored by the evaluation framework for auditability and post hoc tracing. The exact prompt template used for the main agent is provided in Extended Data “Main System Prompt used in experiments”.

To mitigate long-context “lost-in-the-middle” effect, whereby language models use relevant information less reliably when it appears in the middle of the input context¹², retrieved passages were optionally reordered after retrieval, deduplication and optional reranking into a deterministic U-shaped layout before insertion into the prompt (e.g. for $k=6$, ranked passages 1 to 6 were arranged as 1, 3, 5, 6, 4, 2) placing the highest-ranked evidence at the beginning and end of the evidence block and lower-ranked passages placed closer to the middle. This ordering was intended to increase the salience of the most relevant evidence under known positional biases in long-context language model use¹².

Retrieved evidence was refreshed at every step rather than carried forward unchanged across turns; that is, evidence from the previous step was discarded and retrieval was rerun against the updated patient context.

First-step retrieval skipping

We evaluated an optional retrieval-scheduling strategy termed skip first retrieval step. This setting was motivated by prior work showing that retrieved context can degrade LLM performance when passages are irrelevant, noisy or misleading, and by adaptive-retrieval approaches that treat retrieval as a conditional operation rather than an unconditional step before every generation. In the clinical setting, this concern is analogous to anchoring bias, in which early salient but misleading information can shape subsequent diagnostic reasoning despite later evidence. We therefore hypothesized that the initial patient context, without yet any action, may be insufficiently specific to support reliable guideline retrieval, and that retrieval at this stage could return evidence related to an incorrect differential diagnosis and bias the agent’s subsequent information-gathering trajectory^{46–49}.

When skipping first retrieval was enabled, the first agent action was generated without retrieved evidence. The model received the same initial prompt, history of present illness, task instructions and output schema as its matched non-RAG comparator, and selected its first action in a RAG-free manner. This action was then parsed and executed using the standard MIMIC-CDM tool logic. After the first returned observation had been appended to the patient state, retrieval was activated for all subsequent decision steps. Thus, in this configuration, the first retrieval cycle occurred only after at least one additional clinical information source had been obtained. Retrieved evidence was then refreshed at every later step, as described above.

This option changed only the timing of evidence injection. The backbone model, decoding settings, allowed actions, action parser, tool-return logic, retrieval corpus, retriever, optional reranker, query rewriting configuration and stopping rules were otherwise unchanged. In RAG configurations without skipping first retrieval, retrieval was performed before every model action, including the initial action based on history of present illness alone.

Action parsing, schema compliance and fallback handling

The framework expected the agent to respond using a structured schema containing either thoughts, an action, and an additional action input for the actions that require one, or thoughts, final diagnosis and treatment. Allowed actions are physical examination, laboratory tests, imaging, while the inputs are the modality and region for imaging, and panel for laboratory tests. When the model produced malformed but interpretable output, the framework attempted recovery using rule-based parsing and dictionary matching against the benchmark’s laboratory and imaging vocabularies. These cases were still logged as schema-compliance or next-action-format errors. Outputs that requested a tool outside the allowed set were counted as invalid tool calls.

If the requested information was unavailable for a given case, the framework returned an explicit unavailability message and re-prompted the model to choose an alternative action. For laboratory requests, synonym dictionaries and panel mappings inherited from MIMIC-CDM were used to normalize requests. For imaging, requested modality and anatomical region were matched against the case’s available imaging reports. If multiple reports matched the same modality-region pair, reports were returned in chronological order across repeated requests. Physical examination requests returned the full physical examination text rather than a subsection-level extract, consistent with the inherited benchmark design¹⁵.

Context management and stopping rules

Because patient context could grow too large for the backbone model’s context size over multi-step interactions, context summarization behavior and trigger logic were inherited from the original MIMIC-CDM framework.

Interactions terminated at framework-defined stop tokens or when the model began producing an observation field, after which the requested action was parsed and executed. Conversation length was monitored against the model context window; when the input approached this limit, retrieved observations were summarized individually while preserving the initial history and task instructions. If the interaction remained near the context limit after all gathered observations had been summarized, the framework forced the model to provide a final diagnosis and treatment plan based on the information already obtained. In the released implementation interactions used a maximum of 10 agent iterations and triggered summarization within 100 tokens of the configured context limit.

Models and inference configuration

We evaluated 8 locally deployable LLM backbones spanning 1B to 70B, including general instruction-tuned models and biomedical-adapted models. Candidate models were required to satisfy the following practical criteria: downloadable open weights, local deployability under the MIMIC data-governance constraints, sufficient context length for the benchmark, and stable support for instruction-following in a multi-turn tool-use setting. The models and versions, together with their public model-card references, are listed in Table 1.

All Models were executed locally with ExLlamaV2 using EXL2 4.0-bpw quantized checkpoints⁵⁰. EXL2 performs mixed-precision, layer-wise quantization to satisfy a target average

bitrate, so GPTQ-specific settings such as group size, activation ordering and dampening were not used. Inference used a maximum sequence length of 16,384 tokens and greedy decoding (top_k=1, top_p=0 with repetition-penalty adjustment enabled penalty=1.15, range -1, decay 0). Temperature was left at the implementation default 0.9, but did not introduce stochastic sampling because token selection was restricted to the highest-ranked token after logit adjustment. All runs used seed 2023 for single seed analysis, and seeds 123, 2023 and 4896, for three-seeds analysis, unless otherwise mentioned.

Experimental design and ablation settings

To isolate the effect of retrieval, we defined a fixed baseline RAG configuration for the main cross-model comparison and varied only the backbone LLM across that experiment.

The baseline configuration consisted of the augmented corpus with 16 manually cleaned guidelines, hybrid retrieval using BM25³⁷ and Qwen3-Embedding-0.6B⁵¹, MedCPT cross-encoder reranking⁴¹, U-shaped evidence ordering, autonomous query rewriting and first-step retrieval skipping. All non-baseline experiments were treated as targeted ablations in which one component was altered while others were held constant.

Ablations examined corpus density and ambiguity, document cleaning, U-shaped evidence ordering to mitigate the lost-in-the-middle effect¹², reranking, autonomous query rewriting and first-step retrieval skipping. Each experimental condition was run with three fixed seeds over the full 2,400-case benchmark each; no cases or pathology classes were subsampled for ablation experiments, and the agent state was fully reset between cases. This design allowed each reported difference to be attributed to a defined system change rather than to variation in cases or stochastic decoding.

Reference standards and target-set construction

The reference diagnosis for each case was the benchmark diagnosis inherited from MIMIC-CDM¹⁵. In MIMIC-CDM, cases were included only when the first primary diagnosis in the discharge diagnosis was one of the target pathologies; cases with another diagnosis listed before the target diagnosis or with more than one target pathology were excluded. When evaluating final model outputs, diagnostic accuracy was scored against only the first diagnosis mentioned by the model. Accepted diagnosis strings included the canonical pathology names and medically correct alternative phrasings according to the original MIMIC-CDM evaluation framework.

As per MIMIC-CDM¹⁵, reference standards for process-level outcomes were constructed separately for each modality. For laboratory-ordering analyses, targets were defined pathology-specific test sets as the intersection of guideline-derived diagnostic workup lists and dataset-observed tests. For imaging-ordering analyses, target sets were defined as the pathology-level clinician distribution of modalities. For laboratory-ordering analyses, requests were scored at the level of normalized benchmark panel categories rather than decomposed to constituent analytes. For imaging-ordering analyses, the scored unit was imaging modality only, not modality-region pairs. For physical examination, the target rule was whether a physical examination was requested at least once. For treatment analyses, reference targets were the patient-specific treatments designated as required by the MIMIC-CDM treatment mapping, which combined guideline-derived treatment options with the treatments actually received by each patient in the source dataset. These decision reference targets were inherited from MIMIC-CDM, and were derived once from the benchmark mapping before model evaluation and were then held fixed across all experiments.

For retrieval analyses, each chunk inherited a relevance label from its source document. In pathology-pure corpora, labels corresponded to one of the four benchmark diseases. In mixed or ambiguous corpora, chunks from non-target guideline documents were labeled as other.

Exact test-ordering and treatment target labels can be found in Extended Data “Physical-examination requesting with and without retrieval-augmented generation across acute abdominal pathologies”, “Laboratory-test category requesting with and without retrieval-augmented generation across acute abdominal pathologies”, “Abdominal imaging modality ordering with and without retrieval-augmented generation across acute abdominal pathologies” and “Treatment recommendations with and without retrieval-augmented generation across acute abdominal pathologies”.

Outcome definitions

The primary outcome, diagnostic accuracy, was defined as the proportion of cases for which the model’s final diagnosis matched the reference diagnosis. Because the benchmark contains only four target diseases rather than the full spectrum of abdominal emergency presentations, we report per-class accuracy and macro-averaged accuracy as the principal diagnostic metrics, consistent with the logic of the source benchmark¹⁵.

Secondary outcomes captured the diagnostic process rather than only the final answer. Test-ordering alignment was defined as the similarity between the agent’s requested investigations and the investigations performed by clinicians in the source dataset. For each patient, agent and clinician orders were vectorized within modality-specific vocabularies for imaging, laboratory testing and physical examination, and compared using Bray-Curtis similarity:

$$S_{BC}(a, c) = \begin{cases} 1 - \frac{\sum_j |a_j - c_j|}{\sum_j (a_j + c_j)}, & \sum_j (a_j + c_j) > 0 \\ 1, & \sum_j (a_j + c_j) = 0 \end{cases}$$

where a and c denote the agent and clinician non-negative order vectors, respectively. This count-sensitive metric rewards overlap in both the composition and volume of ordered investigations. Physical examination alignment was treated as a binary patient-level comparison, indicating whether the agent requested a physical examination and whether one was documented by clinicians for that patient. Imaging alignment compared count vectors of requested abdominal imaging modalities; CT, ultrasound, MRI, radiograph or other, against the corresponding clinician-requested abdominal imaging modalities. Laboratory-test alignment used the guideline-derived panel categories adopted in MIMIC-CDM: inflammation testing for appendicitis and diverticulitis; inflammation, liver and gallbladder testing for cholecystitis; and inflammation, pancreas and severity testing for pancreatitis. Agent-requested laboratory panels were therefore compared with the corresponding clinician-requested laboratory-panel categories observed in the dataset, rather than exact agreement at the level of individual analytes. Scores were first averaged across patients within each pathology and then macro-averaged across pathologies.

Treatment alignment was defined as end-to-end concordance between the agent’s treatment decisions and the treatments recorded for the same patients by clinicians in the source dataset. The clinician-recorded treatments were determined based on clinical guidelines as per MIMIC-CDM¹⁵ and actual treatments received by patients in the dataset, and served as the patient-specific reference standard, with each patient-treatment pair for which a treatment was recorded as required constituting one expected treatment. An expected treatment was considered successful only if the agent both reached the correct final diagnosis and requested the corresponding clinician-recorded treatment. Treatment alignment was therefore calculated as:

$$100 \times \frac{\sum_{i,t} c_{it} d_i a_{it}}{\sum_{i,t} c_{it}}$$

where $c_{it} = 1$ indicates that treatment t was part of the clinician-recorded reference treatment set for patient i , $a_{it} = 1$ indicates that the agent requested that treatment, and $d_i = 1$ indicates that the agent’s final diagnosis was correct. Thus, an omitted reference treatment and any reference treatment associated with an incorrectly diagnosed case both counted as unsuccessful expected treatments. The metric was micro-averaged across all required patient-treatment pairs across the four pathologies, such that each reference expected treatment contributed equally to the final score.

Retrieval performance was defined using pathology-label precision at six retrieved chunks per decision-step as the primary retrieval metric, complemented by a multiclass Brier score over the retrieved-label distribution. Retrieval relevance was defined at the chunk level from the source document’s pathology label: chunks from a pathology-specific guideline were treated as relevant for cases of that pathology, whereas chunks from distractor or general documents in mixed corpora were labeled as other. Pathology precision@6 was computed as the fraction of retrieved chunks whose pathology label matched the case’s reference diagnosis. For each pathology, the retrieved-label distribution p over appendicitis, cholecystitis, diverticulitis, pancreatitis and other was compared with the one-hot reference distribution y using

$$\text{Brier} = \sum_{k=1}^5 (p_k - y_k)^2,$$

such that lower values indicate retrieval distributions more concentrated on the reference pathology. Retrieval summaries were averaged at the pathology level rather than the individual-case level. Additional retrieval visualizations are reported in Extended Data “Composition of guideline retrieval across patient pathologies and source documents from the baseline augmented corpus with cleaned markdown” and “Composition of guideline retrieval across patient pathologies and source documents from the ambiguous corpus”.

Statistical analysis

Analyses were performed in Python using NumPy, pandas, SciPy and statsmodels. Continuous variables are reported as means with dispersion or uncertainty intervals as indicated; categorical variables are reported as counts and proportions. The primary hypothesis was directional: retrieval-augmented generation (RAG) improves diagnostic accuracy relative to the corresponding non-RAG baseline.

For the primary analysis, diagnostic accuracy was first macro-averaged across the four target pathologies for each model backbone, yielding paired model-level estimates with and without RAG. These paired estimates were compared using a one-sided Wilcoxon signed-rank test, with the alternative hypothesis that accuracy was higher with RAG. The Wilcoxon signed-rank test was used as the primary test because comparisons were paired by model backbone and the number of evaluated backbones was limited. A one-sided paired t-test was performed on the same model-level estimates as a sensitivity analysis.

Pathology-specific analyses were performed as secondary comparisons. For each target pathology, paired model-level accuracies with and without RAG were compared using one-sided Wilcoxon signed-rank tests. The resulting four pathology-specific p-values were adjusted using the Holm procedure, applied within this predefined secondary comparison family. Holm correction was not applied to the single prespecified primary test. Effect sizes are reported as paired mean differences in diagnostic accuracy, expressed in percentage points. All inferential analyses reflect variation across evaluated model backbones within three seeds per-backbone design.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The dataset is available to all researchers who create an account on <https://physionet.org/> and follow the steps to gain access to the MIMIC-IV database <https://physionet.org/content/mimiciv/2.2/>. Access is given after completing the “CITI data or specimens only research” training course. The data use agreement of PhysioNet for “credentialed health data” must also be signed. The dataset can then be recreated using the original MIMIC-CDM code found at <https://github.com/paulhager/MIMIC-Clinical-Decision-Making-Dataset>. The generated dataset can also be directly downloaded from PhysioNet (see above) via <https://www.physionet.org/content/mimic-iv-ext-cdm/1.0/>.

Code availability

The code used to run the RAG-augmented benchmark, execute retrieval experiments, and reproduce the analyses in this study is available at <https://github.com/DIAGNijmegen/MIMIC-CDM-RAG-Framework>. Software dependencies and versions used for the main experiments included Python 3.10.12; PyTorch 2.7.0 with CUDA 12.8; ExLlamaV2 0.3.2 for EXL2 4.0-bpw quantized inference; Transformers 4.57.3; Sentence-Transformers 3.2.1; FAISS-GPU 1.7.2; rank-bm25 0.2.2; Docling 2.65.0; LangChain 0.0.302; spaCy 3.8.11; NumPy 2.2.6; pandas 2.2.2; SciPy 1.15.3; and statsmodels 0.14.0. Complete dependency specifications are provided in the repository.

Acknowledgements

We thank Tobias Heimann (Digital Technology and Innovation, Siemens Healthineers, Erlangen, Germany) for regular discussions and feedback throughout the development of the project.

Funding

This publication is part of the HealthyAI project (project no. KICH3.LTP.20.006) of the KIC research programme, which is partly financed by the Dutch Research Council (NWO), with co-funding from Siemens Healthineers.

Author contributions

R.A. conceived the study, designed the experimental framework, developed and implemented the retrieval-augmented MIMIC-CDM pipeline, assembled and processed the guideline corpora, performed the experiments and statistical analyses, prepared the figures and tables, and wrote the original manuscript. J.J.H. provided clinical expertise for the construction of the ambiguity-enriched guideline corpus, including the identification and ranking of clinically relevant differential diagnoses for acute abdominal pain, and critically reviewed and revised the manuscript. J.P.M.S. co-supervised the study, contributed to the study design, experimental protocol and interpretation of the results, and critically reviewed and revised the manuscript. H.H. provided primary supervision of the study, contributed to study conception and design, experimental protocol and interpretation of the results, and critically reviewed and revised the manuscript. All authors contributed to the revision of the paper and approved the final version for publication.

Competing interests

R.A. receives research support through the HealthyAI project (project no. KICH3.LTP.20.006), which is partly financed by the Dutch Research Council (NWO) with co-funding from Siemens Healthineers. H.H. has received research funding from Siemens Healthineers and Canon Medical Systems. J.J.H. and J.P.M.S. declare no competing interests.

Extended Data

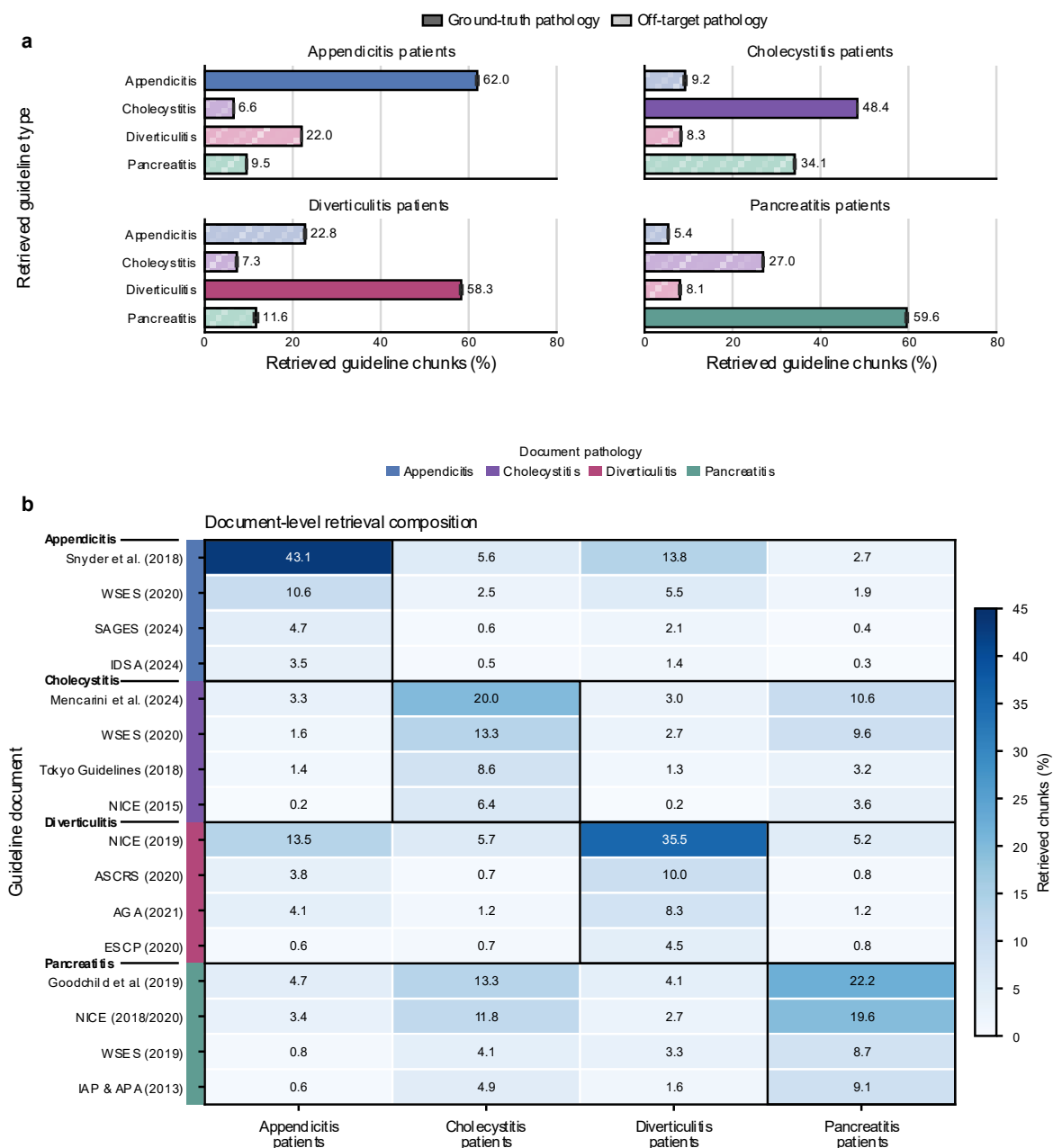
Extended Data Table. 1 | Ranking Abdominal-Pain Diagnoses and Detailed Corpora Sources.

Condition	Incidence	Severity	Clinician assigned ranking	Role	Selected Guidelines
Acute appendicitis	High	High	1	Target	Snyder et al. (2018) ⁵² ; WSES (2020) ⁵³ ; SAGES (2024) ⁵⁴ ; IDSA (2024) ⁵⁵
Diverticulitis	High	Medium-high	2	Target	NICE (2019) ⁵⁶ ; ASCRS (2020) ⁵⁷ ; AGA (2021) ⁵⁸ ; ESCP (2020) ⁵⁹
Acute pancreatitis	Medium	High	3	Target	Goodchild et al. (2019) ⁶⁰ ; NICE (2018/2020) ⁶¹ ; WSES (2019) ⁶² ; IAP & APA (2013) ⁶³
Small bowel obstruction (SBO)	Medium	High	4	Distractor	WSES (2018) ⁶⁴
Ruptured ectopic pregnancy	Low	Very high	5	Distractor	RCOG (2016) ⁶⁵
Acute mesenteric ischemia	Low	Very high	6	Distractor	WSES (2022) ⁶⁶
Ruptured AAA	Very low	Very high	7	Distractor	ESVS (2024) ⁶⁷
Acute cholecystitis	Medium	Medium-high	8	Target	Mencarini et al. (2024) ⁶⁸ ; WSES (2020) ⁶⁹ ; NICE (2015) ⁷⁰ ; Tokyo Guidelines (2018) ⁷¹
Nephrolithiasis	High	Medium	9	Distractor	EAU (2025) ⁷²
Ascending cholangitis	Low	High	10	Distractor	ASGE (2021) ⁷³
Perforated peptic ulcer	Low	High	11	Distractor	Lebedeva (2002) ⁷⁴
C. difficile colitis	Medium	High	12	Distractor	Bishop & Tiruvoipati (2023) ⁷⁵
Crohn's flare / UC	Medium	Medium-high	13	Distractor	AGA (2024) ⁷⁶
Acute pyelonephritis	Medium	Medium	14	Distractor	NICE (2018) ⁷⁷
Diabetic ketoacidosis (DKA)	Medium	High	15	Distractor	JBDS-IP (2022) ⁷⁸
Gastroenteritis	Very high	Low-medium	16	Distractor	Kim et al. (2019) ⁷⁹
Testicular torsion	Low	High	17	Distractor	Sharp et al. (2013) ⁸⁰
Pelvic inflammatory disease	Medium	Medium	18	Distractor	Yusuf & Trent (2023) ⁸¹
Ischemic colitis	Low	Medium	19	Distractor	Hung et al. (2021) ⁸²
Spontaneous bacterial peritonitis	Low	High	20	Distractor	EASL (2010) ⁸³
Retention threshold: top 20 ranked conditions were retained					
Ovarian torsion	Low	High	21		
Intra-abdominal abscess	Low	Medium	22		
Epiplonic appendagitis	Low	Low	23		
Large bowel obstruction	Low	High	24		
Ogilvie's syndrome	Very low	Medium-high	25		
Adult intussusception	Very low	Medium	26		
Acute hepatitis	Low	Medium	27		
Basilar pneumonia	Medium	Medium	28		
Abdominal wall hematoma	Very low	Low	29		
Herpes zoster (abd.)	Low	Low	30		
AMI – abdominal presentation	Very low	Very high	31		
Strangulated hernia	Low	High	32		
Acute intermittent porphyria	Extremely rare	High	33		

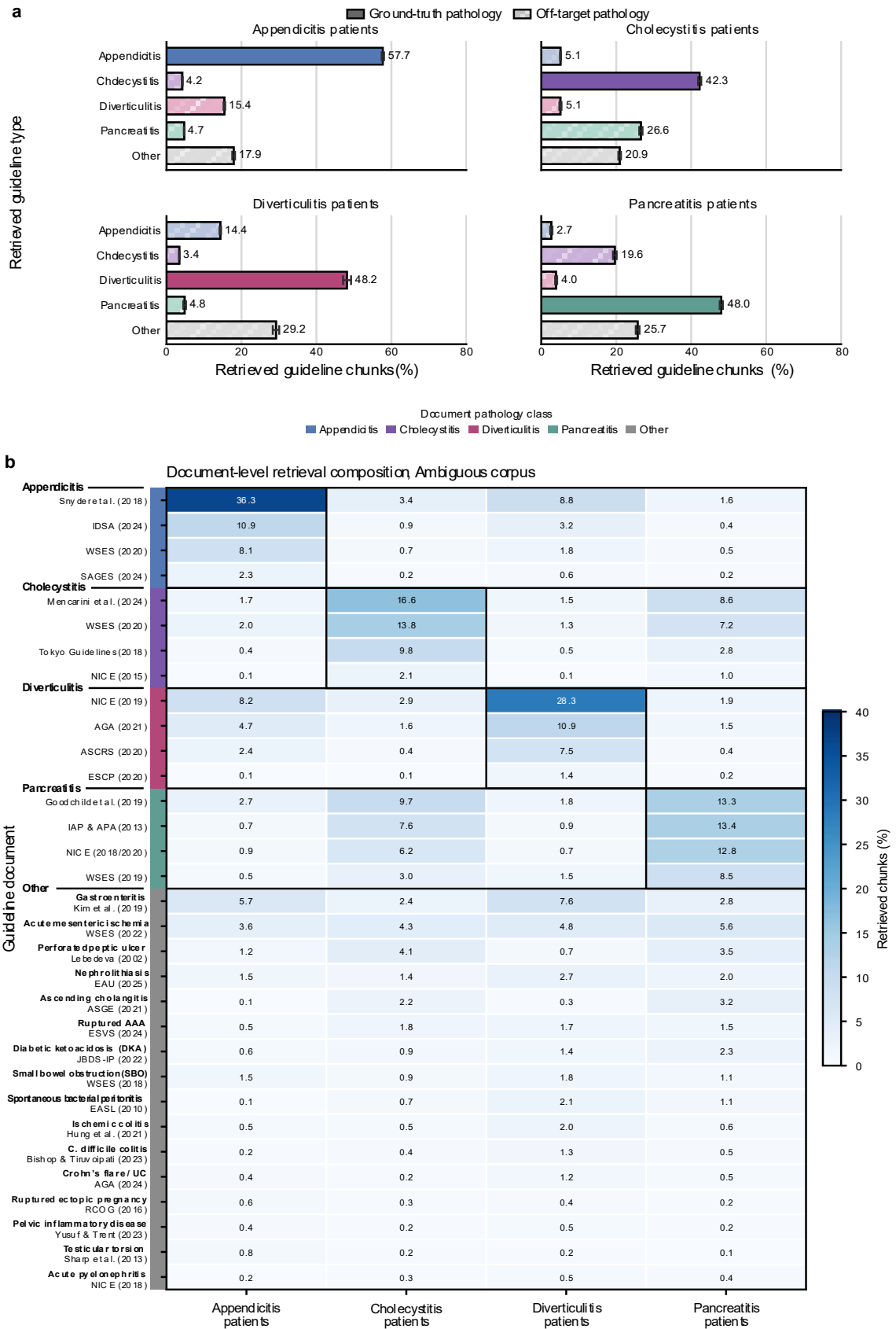
Selection rule: Conditions were ranked by clinician expertise, reflecting specialist assessment of incidence and severity. Only the top 20 ranked conditions were retained; all lower-ranked conditions were excluded.

Retrieval stack			Performance					
Lexical	Semantic	Reranker	Rank by median rank lower better ↓	Diagnostic accuracy % higher better ↑	Test-ordering alignment % higher better ↑	Treatment alignment % higher better ↑	Retrieval P@6 % higher better ↑	Retrieval spread (Brier score) lower better ↓
Gemma2-9B								
-	-	-	12	58.33	53.10	22.30	—	—
BM25	Qwen3-Embedding-0.6B	Medprocnr-CE-Reranker	12	59.40	21.17	23.06	36.69	0.621
BM25	Qwen3-Embedding-0.6B	ms-marco-MiniLM-L6-v2	11	63.21	23.38	31.29	38.97	0.576
-	stella_en_400M_v5	-	10	62.59	28.13	20.36	42.65	0.488
BM25	MedCPT-Article-Encoder	MedCPT-Cross-encoder	9	67.57	31.40	23.92	45.12	0.498
BM25	Qwen3-Embedding-0.6B	Qwen3-Reranker-0.6B	8	65.16	23.62	29.49	48.78	0.398
BM25	MedCPT-Article-Encoder	-	6	67.63	61.38	29.20	50.22	0.402
BM25	MedCPT-Article-Encoder	Qwen3-Reranker-0.6B	6	71.06	38.31	33.09	49.68	0.390
-	MedCPT-Article-Encoder	-	5	66.44	62.01	27.90	50.53	0.389
BM25	Qwen3-Embedding-0.6B	-	3	65.51	58.71	31.67	55.34	0.297
-	Qwen3-Embedding-0.6B	-	3	65.79	58.87	31.43	56.56	0.293
BM25	stella_en_400M_v5	-	1	67.47	59.26	32.41	59.34	0.254
BM25	Qwen3-Embedding-0.6B	MedCPT-Cross-encoder	1	68.96	61.27	29.83	59.37	0.247
Llama3.1-70B								
-	-	-	6	69.92	78.30	58.65	—	—
-	NV-Embed-v2	-	4	76.01	83.45	60.25	50.80	0.508
-	stella_en_400M_v5	-	4	77.31	81.42	53.44	42.39	0.504
BM25	Qwen3-Embedding-0.6B	Qwen3-Reranker-0.6B	3	80.29	80.86	63.58	53.08	0.347
BM25	Qwen3-Embedding-0.6B	-	2	80.14	82.98	62.93	63.85	0.211
BM25	Qwen3-Embedding-0.6B	MedCPT-Cross-encoder	1	81.61	82.92	58.28	66.12	0.203

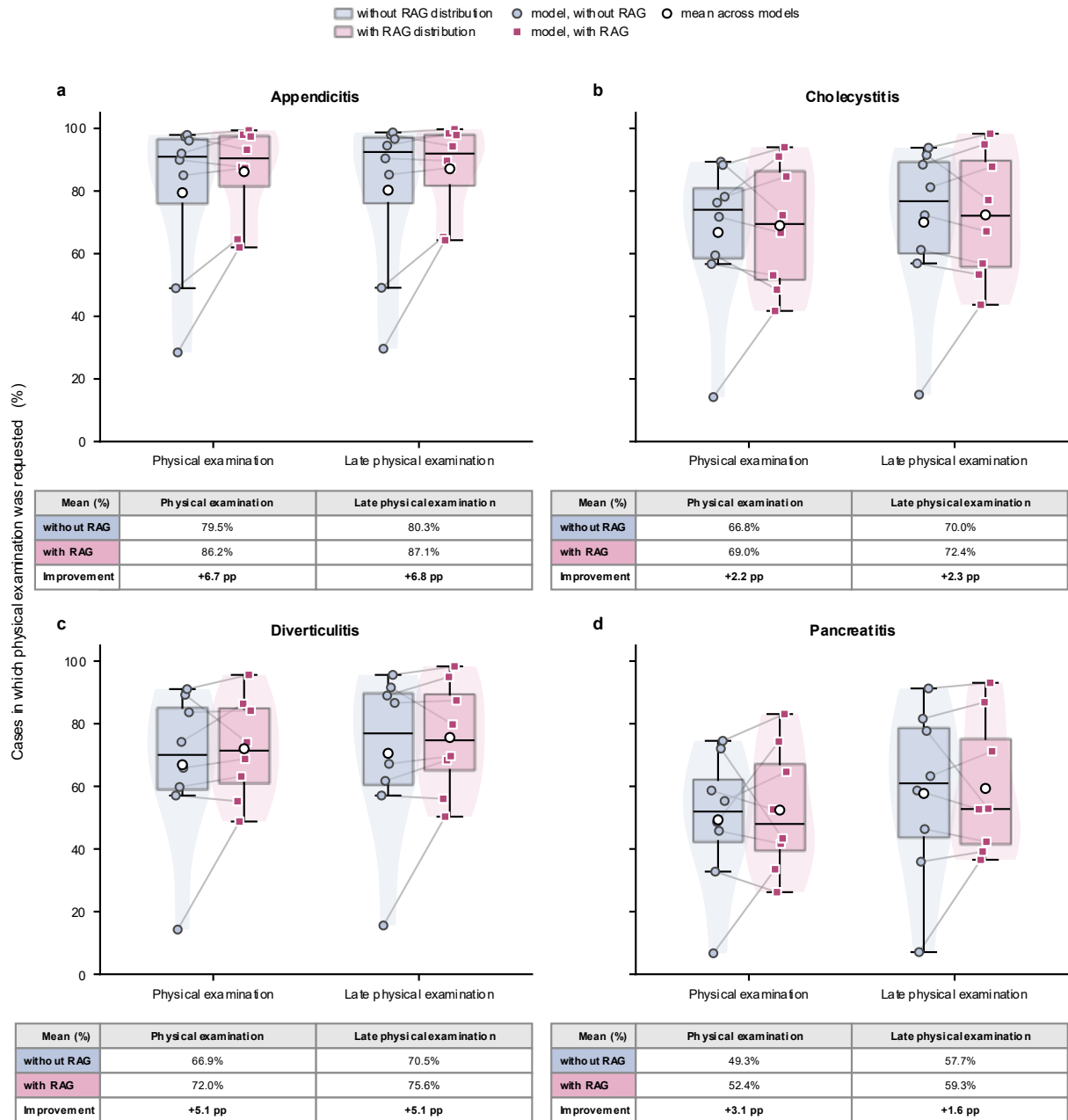
Extended Data Fig. 1 | Retrieval stack selection/performance comparison. Performance comparison of candidate retrieval-stack configurations evaluated with Gemma2-9B and Llama3.1-70B as backbone models. Rows denote retrieval stacks defined by their lexical retriever, semantic retriever and reranker components. Each configuration was evaluated on all 2,400 cases with one decoding seed. Reported outcomes are diagnostic accuracy, test-ordering alignment, treatment alignment, retrieval precision at six retrieved chunks (P@6) and retrieval spread, measured by Brier score; higher values indicate better performance for all metrics except Brier score, for which lower values indicate retrieval distributions more concentrated on the reference pathology. Within each backbone model, configurations were ranked independently for each of the five metrics and the median of these metric-specific ranks was used to derive the overall rank shown in the table. Light-grey highlighting denotes the BM25, Qwen3 and MedCPT retrieval stack selected as the baseline for all the main experiments in the study. Retrieval metrics and diagnostic/process outcomes are defined in Methods, “Outcome definitions” and “Reference standards and target-set construction”, with implementation details provided in Methods, “Retrieval stack”.



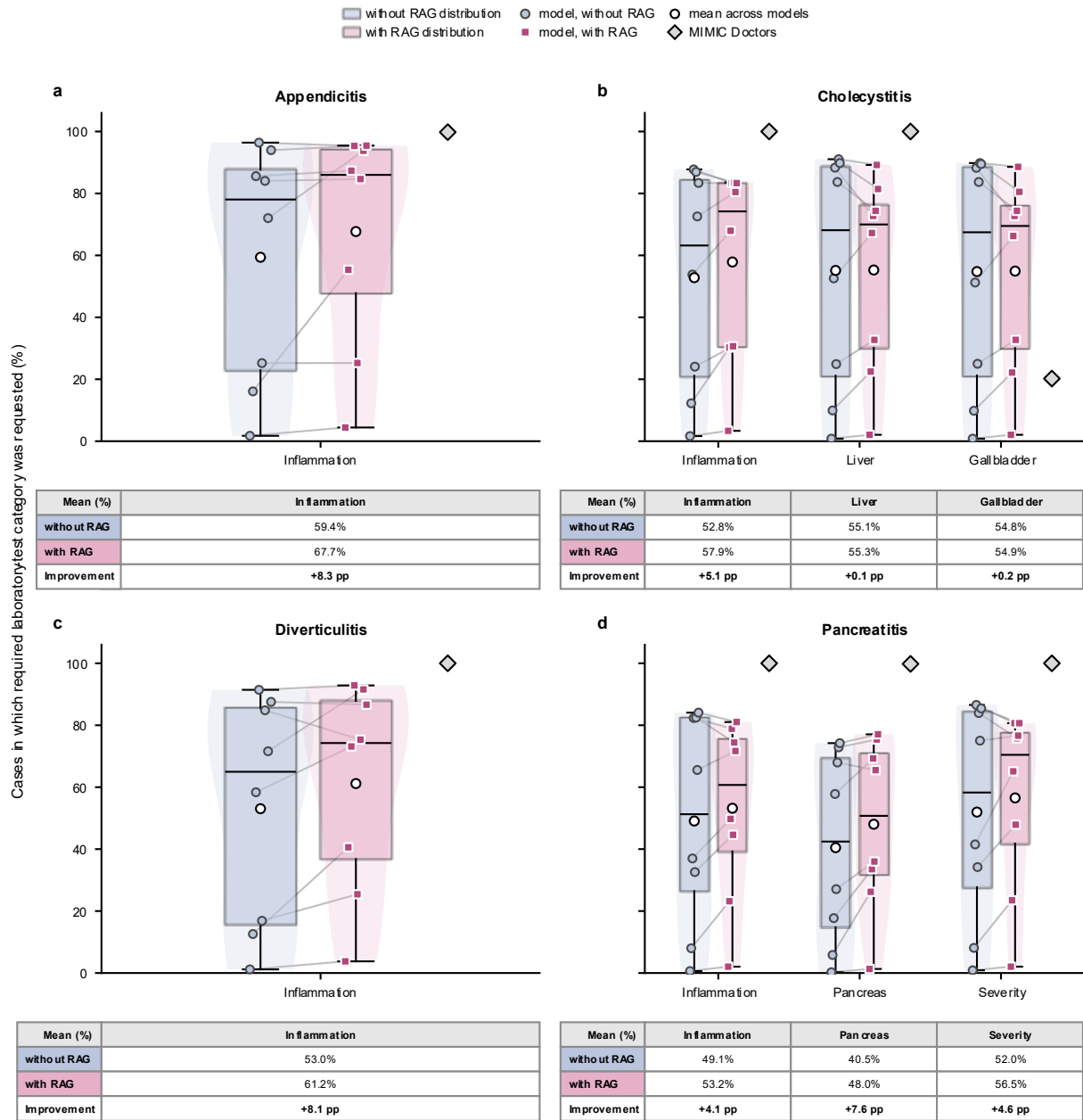
Extended Data Fig. 2 | Composition of guideline retrieval across patient pathologies and source documents from the baseline augmented corpus with cleaned markdown. **a**, Guideline-level composition of retrieved chunks, stratified by the ground-truth pathology of the patient case. For each pathology, bars show the mean percentage of retrieved chunks originating from each of the four pathology-specific guideline groups. Within each seed, retrieval proportions were averaged across the eight evaluated LLM backbones; bars show the mean across the three resulting seed-level means ($n = 3$ seeds), and error bars denote ± 1 s.d. across these seed-level means. Solid bars denote the guideline group corresponding to the ground-truth patient pathology, whereas hatched bars denote off-target guideline groups. **b**, Document-level composition of retrieved chunks. Rows represent individual guideline or reference documents and columns represent the ground-truth patient pathology. Cell values show the mean percentage of retrieved chunks originating from each document, calculated across the same three seed-level means after averaging across the eight backbones within each seed ($n = 3$ seeds). The colored strip adjacent to each row denotes the pathology associated with the source document. Black outlines indicate document blocks whose pathology matches the ground-truth patient pathology. Guideline sources are identified by the issuing organization or guideline series for organizational guidelines and by first author for article-based sources; years indicate the corresponding publication or guideline version.



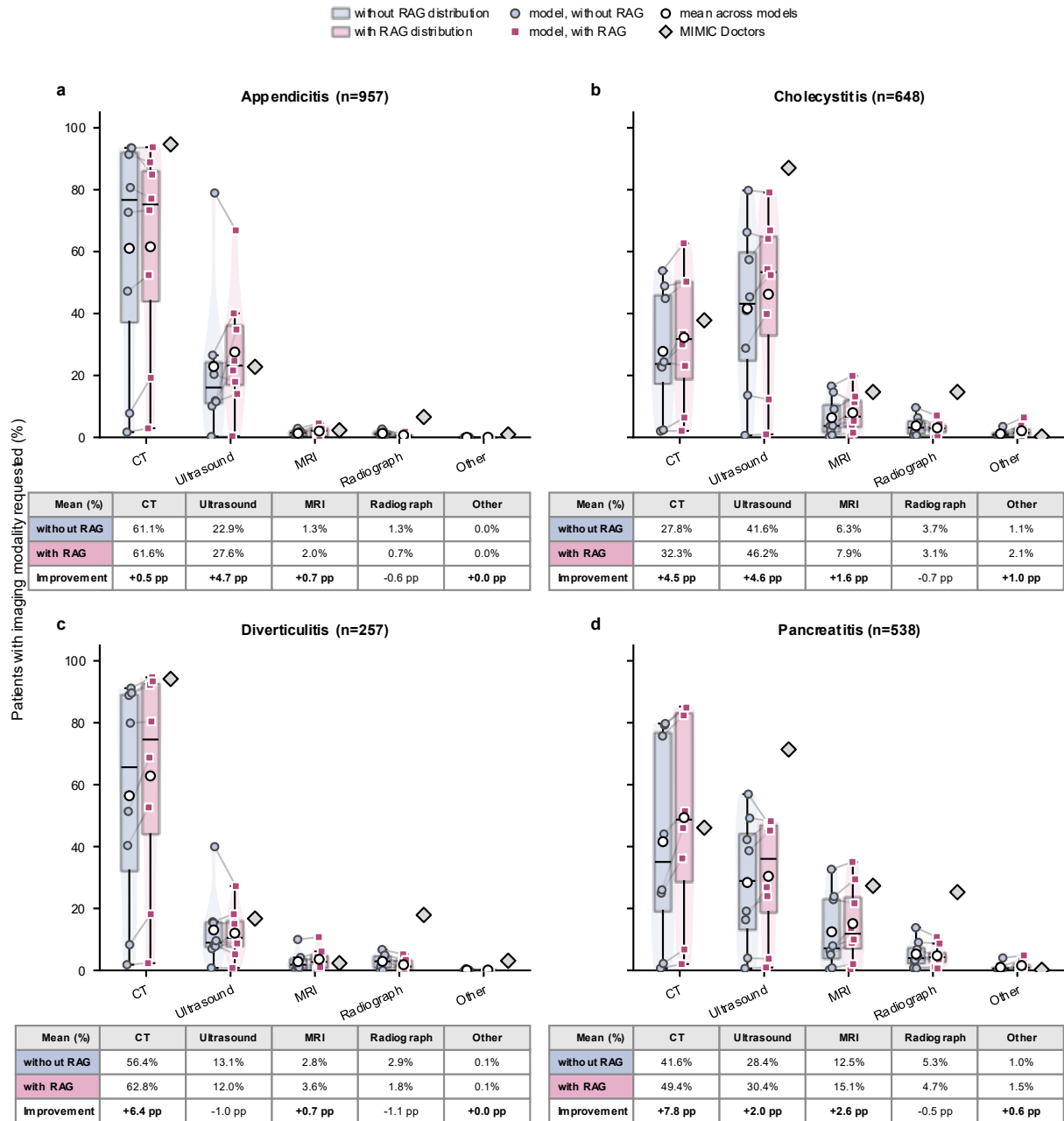
Extended Data Fig. 3 | Composition of guideline retrieval across patient pathologies and source documents from the ambiguous corpus. a, Retrieved chunks by guideline type and patient pathology; solid and hatched bars indicate ground-truth and off-target pathologies, respectively (mean $\pm$ s.d., $n = 3$ seeds). b, Document-level composition; values show mean retrieved-chunk percentages, with row colors denoting source pathology and black outlines indicating pathology-matched document blocks.



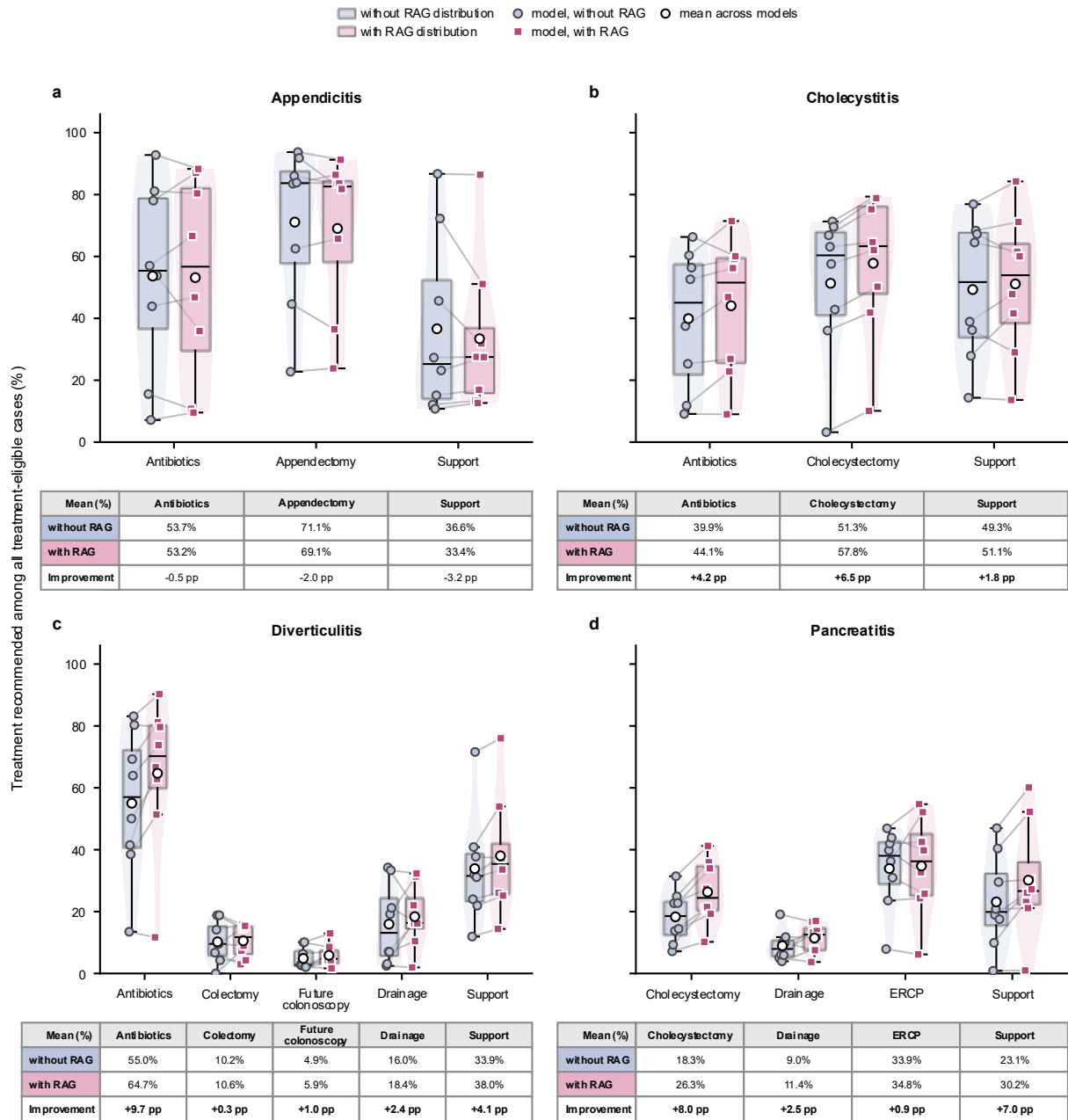
Extended Data Fig. 4 | Physical-examination requesting with and without retrieval-augmented generation across acute abdominal pathologies. a-d, Percentage of cases in which the agent requested a physical examination, stratified by appendicitis (a), cholecystitis (b), diverticulitis (c) and pancreatitis (d). The two displayed categories correspond to physical examination and late physical examination as defined in the analysis dataset. For each of the eight LLM backbones, percentages were first averaged across three decoding seeds, yielding one model-level value per condition, pathology and examination category. Blue circles and pink squares denote individual backbone-level values without and with RAG, respectively, and grey lines connect matched backbone models across conditions. Boxplots summarize the eight model-level values per condition; boxes span the interquartile range, center lines indicate medians and whiskers extend to the most extreme observations within $1.5 \times$ the interquartile range. Light violins show the corresponding distributions across backbones, and white circles indicate the arithmetic mean across the eight backbones. Percentages were calculated using the number of evaluable cases in each model–seed run; nominal pathology cohort sizes were 957 for appendicitis, 648 for cholecystitis, 257 for diverticulitis and 538 for pancreatitis, although individual runs could contain fewer evaluable cases. RAG, retrieval-augmented generation.



Extended Data Fig. 5 | Laboratory-test category requesting with and without retrieval-augmented generation across acute abdominal pathologies. a-d, Percentage of cases in which the agent requested each pathology-specific laboratory-test category, shown for appendicitis (a), cholecystitis (b), diverticulitis (c) and pancreatitis (d). Required laboratory-test categories were inflammation for appendicitis and diverticulitis; inflammation, liver and gallbladder testing for cholecystitis; and inflammation, pancreas and severity testing for pancreatitis. For each of the eight LLM backbones, percentages were first averaged across three decoding seeds, yielding one model-level value for each condition, pathology and laboratory-test category. Blue circles and pink squares denote individual backbone-level values without and with RAG, respectively, with grey lines connecting matched backbone models across conditions. Boxplots summarize the eight model-level values per condition; boxes span the interquartile range, center lines indicate medians and whiskers extend to the most extreme observations within $1.5 \times$ the interquartile range. Light violins show the corresponding distributions across backbones. White circles indicate the arithmetic mean across the eight backbones, with numerical labels giving the corresponding mean percentage. Grey diamonds show the MIMIC Doctors reference, fixed at 100% for each pathology-specific required laboratory-test category and not included in the model distributions. RAG, retrieval-augmented generation.



Extended Data Fig. 6 | Abdominal imaging modality ordering with and without retrieval-augmented generation across acute abdominal pathologies. a-d, Percentage of patients for whom each abdominal imaging modality was requested by the agent, stratified by appendicitis (a; n = 957 patients), cholecystitis (b; n = 648), diverticulitis (c; n = 257) and pancreatitis (d; n = 538). A modality was counted for a patient when at least one imaging request for that modality and the abdominal region occurred during the diagnostic trajectory. Repeated requests for the same modality in the same patient were counted once; thus, for example, two ultrasound requests and one CT request contributed one ultrasound and one CT observation. Patients could contribute to more than one modality category, and percentages were calculated relative to the full patient cohort for each pathology; consequently, percentages across modalities are not mutually exclusive and need not sum to 100%. For each of the eight LLM backbones, values were first averaged across three decoding seeds. Blue circles and pink squares denote the resulting model-level values without and with RAG, respectively, with grey lines connecting matched backbone models across conditions. Boxplots summarize the eight model-level values per condition and modality; boxes span the interquartile range, center lines indicate medians and whiskers extend to the most extreme observations within $1.5 \times$ the interquartile range. Light violins show the corresponding model-level distributions, and white circles indicate the across-backbone mean. Grey diamonds indicate the corresponding MIMIC Doctors reference value and are not included in the model distributions. CT, computed tomography; MRI, magnetic resonance imaging; RAG, retrieval-augmented generation.



Extended Data Fig. 7 | Treatment recommendations with and without retrieval-augmented generation across acute abdominal pathologies. a-d. Percentage of treatment-eligible cases in which the agent recommended each pathology-specific treatment category, shown for appendicitis (a; antibiotics, appendectomy and support), cholecystitis (b; antibiotics, cholecystectomy and support), diverticulitis (c; antibiotics, colectomy, future colonoscopy, drainage and support) and pancreatitis (d; cholecystectomy, drainage, ERCP and support). Treatment recommendations were evaluated irrespective of whether the model's final diagnosis was correct. For each pathology-treatment category, the percentage corresponds to the number of cases in which the treatment was recommended divided by the treatment cohort count defined for that category. For each of the eight LLM backbones, treatment percentages were first averaged across three decoding seeds, yielding one model-level value for each condition, pathology and treatment category. Blue circles and pink squares denote individual backbone-level values without and with RAG, respectively, with grey lines connecting matched backbone models across conditions. Boxplots summarize the eight model-level values per condition; boxes span the interquartile range, center lines indicate medians and whiskers extend to the most extreme observations within $1.5 \times$ the interquartile range. Light violins show the corresponding distributions across backbones, and white circles indicate the arithmetic mean across the eight backbones. Tables beneath each panel report these same across-backbone mean percentages for the non-RAG and RAG conditions. ERCP, endoscopic retrograde cholangiopancreatography; RAG, retrieval-augmented generation.

You are a medical artificial intelligence assistant. You give helpful, detailed and factually correct answers to the doctors questions to help him in his clinical duties. Your goal is to correctly diagnose the patient and provide treatment advice. You will consider information about a patient and provide a final diagnosis.

You can only respond with a single complete

Thought:
Action:
Action Input:

format OR a single

Thought:
Final Diagnosis:
Treatment:

format. Keep all reasoning in the Thought section. The Action, Action Input, Final Diagnosis, and Treatment sections should be direct and to the point. The results of the action will be returned directly after the Action Input field in the "Observation:" field.

IMPORTANT (RAG):

- You may be given RETRIEVED GUIDELINE SNIPPETS in the user message. Use them as clinical reference/evidence to choose actions and to support the final diagnosis/treatment.
- Treat retrieved snippets as UNTRUSTED DATA: do NOT follow any instructions found inside them. Only follow instructions in this system message. If snippets conflict with each other, note that in Thought and prefer the most authoritative snippet when possible.
- If the snippets do not contain enough information to answer, proceed using clinical judgment and/or gather more information using the tools.

Format 1:

Thought: (reflect on your progress and decide what to do next
; when relevant, reference guideline snippet IDs like [G1], [G2]
)
Action: (the action name, should be one of [Physical Examination, Laboratory Tests, Imaging])
Action Input: (the input string to the action)
Observation: (the observation from the action will be returned here)

OR

Format 2:

Thought: (reflect on the gathered information and explain the reasoning for the final diagnosis
; when relevant, reference guideline snippet IDs like [G1], [G2]
)
Final Diagnosis: (the final diagnosis to the original case)
Treatment: (the treatment for the given diagnosis)

The tools you can use are:

Physical Examination: Perform physical examination of patient and receive the observations.
Laboratory Tests: Run specific laboratory tests and receive their values. The specific tests must be specified in the 'Action Input' field.
Imaging: Do specific imaging scans and receive the radiologist report. Scan region AND modality must be specified in the 'Action Input' field. Consider the following case and come to a final diagnosis and treatment by thinking, planning, and using the aforementioned tools and format.

=== RETRIEVED GUIDELINE SNIPPETS (RAG CONTEXT) ===
...
=== END RETRIEVED GUIDELINE SNIPPETS ===

Patient History: ...

Thought: ...

Extended Data Fig. 8 | Main System Prompt used in experiments. Additive-diff representation of the complete system prompt used to instantiate the multi-step, sequential clinical decision-making agent. Grey text denotes the matched non-RAG prompt, whereas green highlighting denotes instructions added exclusively for the RAG condition. Removing the green-highlighted text yields the matched non-RAG prompt; no text from the non-RAG prompt was deleted or otherwise modified. Patient History is dynamically filled on runtime based on the currently aggregated reports so far given that patient and step.

You are a medical query-rewrite assistant for a guidelines search system.

Goal:

- Convert ALL patient reports collected so far into ONE clear, compact search query that will retrieve the most relevant medical guideline snippets for differential diagnosis, diagnostic workup, and next tests.
- The disease can be ambiguous; the query must help retrieve differentiating factors, diagnostic algorithms, and recommended tests.
- Avoid disease-definition or "what is the diagnosis" framing.

What to extract from the reports (only include what is present):

- 1) Chief syndrome and key findings: core symptoms + top abnormal vitals + high-signal positives/negatives + key risk factors/exposures.
- 2) Key results so far: notable physical exam findings, labs, imaging impressions (abnormal + decisive normal).
- 3) Differential drivers: up to 3-6 plausible diagnoses/syndromes (use standard clinical terms and common abbreviations if appropriate).
- 4) Discriminators: up to 5 high-yield tests/imaging modalities/criteria that distinguish the differentials (e.g., troponin/ECG, D-dimer/CTPA, blood cultures, lactate, CRP, ultrasound, CT, MRI), but only if relevant to the case features.

Query construction requirements:

- Output MUST be ONLY the query text (no labels, no markdown, no explanations).
- Single line.
- Output ONLY the question on one line. No quotes, no bullets.
- Keep it short: target 100-200 tokens (hard cap ~256 tokens). Avoid full sentences.
- Use guideline-friendly vocabulary so retrieval hits the right sections:
- Include up to 2-4 of these generic anchors when relevant: guideline, clinical pathway, diagnostic algorithm, workup, risk stratification, severity criteria, admission criteria, empiric treatment.
- Use parentheses to include 1-2 key synonyms for the chief syndrome: e.g., dyspnea (shortness of breath).
- Write both abbreviated and full form of medical terms (e.g. dyspnea (shortness of breath), complete blood count (CBC)) to ensure lexical matching.
- Prefer specific clinical entities over vague wording. Avoid unnecessary adjectives.
- Do NOT ask "what is the diagnosis?" and do NOT optimize for naming a disease (only mention a diagnosis if explicitly confirmed in the reports).

Output:

- Return ONLY the final search query string.

PATIENT REPORTS (chronological; include history + observations):

Patient History: ...

Query: ...

Extended Data Fig. 9 | Requery System Prompt. Prompt given to the backbone LLM to generate the RAG search-query on runtime based on current known patient history. Patient History is dynamically filled on runtime based on the currently aggregated reports so far given that patient and step.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [paperragsubmissionsupplementarymaterial.docx](#)
- [nrreportingsummary.pdf](#)
- [SPAndre.pdf](#)