Characterizing Population Gaps in Clinical Decision-Making: A Guideline-Based Benchmark and Population-Aware Retrieval Analysis in Anesthesiology

Chaiho Shin* CHAIHOYAH@SNU.AC.KR
Interdisciplinary Program of Medical Informatics, College of Medicine
Seoul National University
Seoul, Republic of Korea

Jung-Bin Park* JB4001@SNU.AC.KR
Department of Anesthesiology and Pain Medicine
Seoul National University Hospital
Seoul, Republic of Korea

Kwangsoo Kim† KWANGSOOKIM@SNU.AC.KR
Department of Transdisciplinary Medicine
Seoul National University Hospital
Seoul, Republic of Korea

Hee-Soo Kim† DAMI0605@SNU.AC.KR
Department of Anesthesiology and Pain Medicine
Seoul National University Hospital
Seoul, Republic of Korea

Abstract

Large language models (LLMs) have shown strong performance on medical question answering, yet their ability to support population-specific, action-oriented clinical decision-making remains underexplored. Our analysis of five widely used medical QA benchmarks shows that pediatric anesthesia accounts for only 0.2% of all questions, revealing a critical evaluation gap. To address this, we construct PopAnesQA, a population-aware anesthesiology question-answering benchmark grounded in clinical guidelines, and introduce Prof-RAG, a diagnostic framework for isolating the effects of patient-specific context on retrieval and downstream reasoning. Across diverse LLMs, we observe a consistent pediatric performance gap relative to the general subset, with differences exceeding 15 percentage points even in proprietary models like GPT-4o. We find that incorporating patient profile information leads to highly model-dependent effects, improving performance in some models while degrading it in others. Retrieval trajectory analysis shows that incorporating profile information increases the proportion of population-aligned documents, but does not consistently improve relevance to the clinical question. Overall, our findings highlight that effective clinical retrieval requires not only population awareness, but also careful alignment between patient-specific context and clinical intent.

* These authors contributed equally.
† Corresponding authors.

1. Introduction

Recent advances in large language models (LLMs) have led to remarkable progress in medical question answering and clinical reasoning (Singhal et al., 2025). Along with these advances, benchmark development has moved beyond testing broad medical knowledge toward evaluating specialty-specific competence (Zhao et al., 2023; Holmes et al., 2023; Bhayana, 2024). Nevertheless, existing evaluation literature remains largely concentrated on diagnosis-oriented tasks or assessments of general clinical knowledge, while relatively few studies examine management-oriented clinical decision-making. A recent meta-analysis shows that over half of clinical LLM evaluation tasks focus on diagnosis or basic knowledge, with substantially fewer assessments targeting treatment or management reasoning (Kim and Yoon, 2025). This imbalance limits our understanding of whether LLMs can reliably support action-oriented clinical decisions, where inappropriate recommendations may directly impact patient safety.

This limitation is particularly critical in anesthesiology, where clinical safety depends on selecting appropriate management strategies tailored to patient-specific conditions (Dobson et al., 2025). Unlike diagnosis-centered tasks, anesthesiology practice involves continuous, high-stakes, and context-dependent decision-making under dynamic conditions (Gaba, 1992; Stiegler and Tung, 2014), requiring clinicians to continuously select and adapt management actions as the patient states evolve. This complexity is further amplified by population-specific physiological differences, particularly in pediatric anesthesiology, where management strategies often diverge from adult practice even when clinical presentations are similar (Sury et al., 2015; De Graaff et al., 2021). Failure to account for such differences may lead to inappropriate or unsafe management decisions. Despite these risks, existing benchmarks rarely evaluate whether LLMs can correctly distinguish and apply pediatric-specific management principles in action-oriented scenarios, leaving it unclear whether strong performance on general medical or adult-oriented benchmarks translates to safe clinical decision-making.

To quantify this gap at scale, we analyze the composition of widely used medical benchmarks using LLMs (Table 1). The analysis reveals that anesthesiology-related questions account for only 3.0% of all instances, with pediatric anesthesiology comprising just 0.2%, indicating limited coverage of population-specific perioperative decision-making. Consequently, strong performance on general medical or adult-oriented benchmarks may provide misleading evidence of clinical competence, raising concerns about the safe deployment of such systems in pediatric and other population-sensitive clinical settings. This raises a critical question: *can LLMs accurately perform action-oriented clinical decision-making that is both guideline-consistent and appropriately adapted to patient populations?*

Motivated by these gaps, we introduce PopAnesQA, a population-aware question-answering benchmark for anesthesiology that evaluates action-oriented clinical decision-making across adult, pediatric, and general settings. The benchmark is constructed using an LLM-based pipeline designed to improve consistency and clinical validity, and explicitly tests whether LLMs can identify the intended target population and generate management decisions consistent with patient-specific characteristics and clinical guidelines. In addition, we introduce a profile-aware retrieval diagnostic framework (Prof-RAG) to systematically analyze how query composition affects retrieval dynamics and downstream reasoning outcomes. Prof-

Table 1: Composition of five widely used medical QA benchmarks: MedMCQA (Pal et al., 2022), MedQA (Jin et al., 2021), MMLU (Hendrycks et al., 2020), MedXpertQA (Zuo et al., 2025), and MedBullets (Chen et al., 2025). For MMLU, we use the Professional Medicine and Clinical Knowledge subsets. Columns report counts and within-benchmark percentages for diagnosis or basic-science question type, pediatric target population, anesthesiology relevance, and the intersection of pediatric target population and anesthesiology relevance. Classification and verification details are provided in Section 3.1. Abbreviations: Diag./Basic Sci., Diagnosis or Basic Science; Target Pop., Target Population.

Benchmark	Question Type (Diag./Basic Sci.)	Target Pop. (Pediatric)	Anesthesiology	Pediatric Anesthesiology	Total
MedMCQA	5,659 (92.7%)	452 (7.4%)	230 (3.8%)	5 (0.1%)	6,107
MedQA	911 (71.7%)	242 (19.0%)	20 (1.6%)	2 (0.2%)	1,271
MMLU	402 (75.4%)	60 (11.3%)	27 (5.1%)	0 (0.0%)	533
MedXpertQA	1,348 (55.8%)	446 (18.5%)	42 (1.7%)	10 (0.4%)	2,417
MedBullets	160 (52.0%)	68 (22.1%)	2 (0.7%)	0 (0.0%)	308
Total	**8,480 (79.7%)**	**1,268 (11.9%)**	**321 (3.0%)**	**17 (0.2%)**	**10,636**

RAG enables a structured analysis of when and why retrieval-based systems succeed or fail in clinically sensitive contexts.[1]

Our key findings are summarized as follows:

- **Evaluation gap:** Existing open LLM benchmarks do not adequately evaluate action-oriented decision-making in pediatric and population-specific clinical settings.
- **Population-specific generalization limitations:** LLMs exhibit inconsistent and sometimes unreliable performance when adapting general medical knowledge to population-specific clinical decision-making.
- **Retrieval sensitivity to query composition:** The composition of patient profile and question intent can lead to unstable, population-misaligned, and potentially misleading retrieval outcomes.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work provides three practical insights for deploying LLMs in clinical settings. First, strong general-benchmark performance does not guarantee reliable decisions in population-specific contexts, so evaluation must measure subgroup performance to expose clinically meaningful failures. Second, retrieval is highly sensitive to how patient information enters the query — combining question intent with patient profile trades population relevance against task alignment, producing both gains and degradation. Third, retrieval effectiveness varies across models, giving divergent outcomes under identical configurations and requiring model-specific validation rather than assumed uniform benefit.

1. The benchmark is available at https://huggingface.co/datasets/BMILab/PopAnesQA, and all associated code, including the construction pipeline and evaluation framework, at https://github.com/chaihoyah/PopAnesQA.

2. Related Work

2.1. LLM Evaluation Benchmarks in Medicine

The growing adoption of LLMs in healthcare has led to the development of benchmarks that evaluate clinical knowledge and reasoning. Early benchmarks, such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and medical subsets of MMLU (Hendrycks et al., 2020), primarily assess factual knowledge and exam-style question answering. More recent efforts, including MedXpertQA (Zuo et al., 2025) and MultiCogEval (Zhou et al., 2025), aim to evaluate multi-step clinical reasoning and structured decision-making processes. However, these benchmarks largely focus on diagnostic inference or knowledge synthesis, with limited evaluation of action-oriented clinical decision-making, such as selecting guideline-consistent management strategies under dynamic conditions. Moreover, most benchmarks assume a general patient population and rarely assess whether models can adapt decisions to population-specific contexts, particularly in pediatric settings where inappropriate management may pose significant safety risks. Consequently, it remains unclear whether strong performance on existing benchmarks reflects the ability of LLMs to make safe and context-appropriate clinical decisions in real-world practice.

2.2. Retrieval-Augmented Generation in Clinical Decision Support

Retrieval-augmented generation (RAG) has been widely adopted in clinical settings to enhance the reliability of LLMs by incorporating external medical knowledge, such as biomedical literature, clinical guidelines, and domain-specific databases (Lewis et al., 2020; Xiong et al., 2024a,b; Singhal et al., 2025). Recent studies have further explored the construction of specialty-specific retrieval corpora to improve performance in particular clinical domains, enabling more targeted and knowledge-grounded responses (Luo et al., 2024; Raja et al., 2024; Liu et al., 2025; Wong and Wong, 2026). However, prior work has also shown that retrieval does not consistently improve model performance in clinical settings, and may introduce noise or irrelevant evidence that degrades decision quality (Fensore et al., 2025; Kim et al., 2025; Neha et al., 2025). We hypothesize that the composition of heterogeneous query components—such as patient profiles and clinical questions—introduces structured noise that can systematically affect retrieval behavior. This issue is particularly critical in clinical decision-making, where queries often implicitly combine patient-specific context with task intent, and where misaligned retrieval may lead to population-inappropriate or clinically unsafe recommendations. To better understand this problem, we analyze how query composition influences retrieval outcomes in a population-aware setting.

3. Methods

3.1. Existing Benchmark Composition Analysis

Before constructing our benchmark, we first characterized how existing medical QA benchmarks cover population-specific, action-oriented anesthesiology questions. We classified each question from the benchmarks summarized in Table 1 along three axes: question type, target population, and anesthesiology relevance. Question type was categorized as management, diagnosis, basic science, or other; target population was categorized as pediatric, adult, or

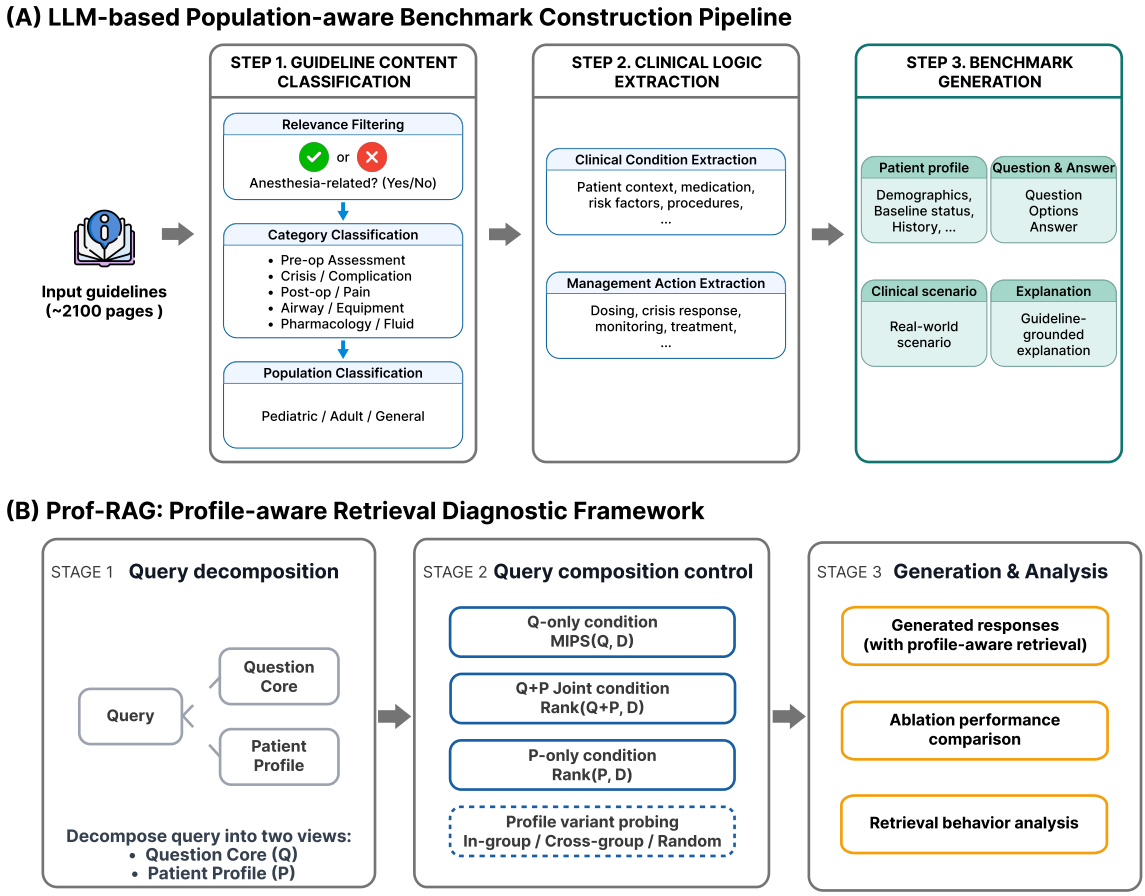

Figure 1: Study overview. (A) LLM-based construction pipeline for PopAnesQA. All stages are performed by a state-of-the-art LLM, yielding structured QA instances with patient profiles, clinical scenarios, and guideline-grounded explanations. (B) Prof-RAG decomposes each query into a question core and a patient profile for controlled query composition. Responses are then analyzed for performance and retrieval trajectory to examine how composition shapes retrieval.

general; and anesthesiology relevance was annotated as a binary label. Classification was performed by majority voting across three LLMs: Gemma-3-27B (Gemma Team, 2025), MedGemma-27B (Sellergren et al., 2025), and Llama-3.3-70B (Grattafiori et al., 2024).[2]

Inter-model agreement across the three LLM classifiers was high for question type (Fleiss' $\kappa = 0.92$; Fleiss, 1971) and target population ($\kappa = 0.81$), and moderate for anesthesiology relevance ($\kappa = 0.61$). To verify the reliability of the LLM-based classification, we further conducted human verification on 100 samples, stratified to ensure balanced representation across question type, target population, and anesthesiology relevance. Two board-certified anesthesiologists independently annotated each instance along the same three axes,

2. Fleiss' κ was computed over all classified questions. For the counts in Table 1, we retained only questions with a majority label, excluding cases with full three-way disagreement (2 for target population, 80 for question type), yielding 10,636 of 10,718 questions.

and disagreements were resolved through consensus. The final expert labels were compared with LLM majority voting, showing near-perfect agreement for target population (Cohen's $\kappa = 0.98$; Cohen, 1960), substantial agreement for question type ($\kappa = 0.78$), and moderate agreement for anesthesiology relevance ($\kappa = 0.59$). Notably, expert reviewers applied stricter criteria for anesthesiology relevance than the LLMs (29 vs. 40 out of 100 instances labeled as relevant), suggesting that the anesthesiology coverage in Table 1 may even be an overestimate. Nevertheless, the conclusion of severe pediatric anesthesiology underrepresentation remains robust.

3.2. PopAnesQA: Guideline-based Population-Aware Benchmark Construction

Clinical guidelines serve as authoritative resources for scenario-specific decision-making in anesthesiology. Our goal is to construct a question-answer benchmark grounded in clinical contexts and corresponding management actions derived from these guidelines. We curated approximately 100 open-access anesthesiology guidelines (Table A2), selected by an anesthesiology specialist to ensure coverage across clinical scenarios and patient populations, especially balancing general and pediatric contexts. The collected documents were converted into raw text using the Upstage Document Parse API and preprocessed by removing redundant whitespace. To filter out non-informative fragments, we retained only text segments exceeding 100 characters. Based on this processed corpus, we design a three-step construction pipeline to ensure consistency and guideline-grounded validity, as illustrated in Figure 1 (A). All three steps of the pipeline are performed with Gemini-3-Flash, selected through a small-scale qualitative comparison by a domain specialist (Appendix A).

Step 1. Guideline Content Classification
The first step of benchmark construction involves filtering out irrelevant content and classifying each guideline segment by target population and clinical category. Due to OCR artifacts and the presence of non-informative sections (e.g., references), we first identify and remove segments that are not relevant to anesthesiology. For the remaining segments, we perform category classification into five predefined groups: Pre-operative Assessment, Crisis & Complication, Post-operative Care & Pain, Airway & Equipment, and Pharmacology & Fluid Management. In parallel, each segment is assigned a target population label among Pediatric, Adult, and General. General population denotes guideline actions that apply comparably to pediatric and adult patients; we keep this label separate from Adult rather than collapsing population-agnostic guidance into an adult-specific category, which would otherwise mislabel broadly applicable recommendations. Both relevance filtering and classification are performed simultaneously using a unified prompt (Prompt A1).

Step 2. Clinical Logic Extraction
The next step focuses on extracting structured clinical action logic from guideline content. Clinical guidelines primarily provide action-oriented recommendations under specific clinical conditions. To capture this structure, we extract condition–action pairs that represent executable clinical rules. The LLM is constrained to perform strict extraction without introducing external knowledge or inference, ensuring fidelity to the source material.

Each extracted rule is represented in a structured format consisting of (1) a condition base describing the patient context, (2) triggers specifying criteria for action, (3) the recommended action, (4) the source fragment from which the rule is derived, and (5) optional notes for clarification. Extraction is performed within the target population and clinical category identified in Step 1, ensuring consistency across the pipeline (Prompt A2). This pipeline ensures that all QA instances remain grounded in guideline-derived clinical logic.

Step 3. Question–Answer Pair Generation
The final step generates question–answer pairs based on the structured clinical logic extracted in the previous steps. Each question and its corresponding answer are grounded in the extracted condition–action rules, ensuring alignment with guideline-based recommendations. To ensure factual accuracy, we incorporate the original guideline text as a validation reference during generation, instructing the LLM to verify key details against the source.

In addition, we generate corresponding patient profiles and clinical scenarios conditioned on the extracted condition base and trigger criteria. Each instance is represented in a structured format consisting of: question type, target population, clinical category, patient profile, clinical scenario, question, answer options, correct answer, and explanation (Prompt A3). Overall, the pipeline yielded PopAnesQA, a benchmark of 623 instances comprising 314 pediatric, 159 general, and 150 adult cases, each paired with corresponding patient profiles and clinical scenarios.

3.3. Expert Quality Evaluation Protocol

We conducted a human evaluation on 100 randomly sampled QA instances from PopAnesQA (50 pediatric, 25 adult, and 25 general), ensuring diversity across guideline sources. Two board-certified anesthesiologists independently evaluated each instance using a predefined protocol. Four criteria—Guideline Consistency, Medical Validity, One-Best-Answer Clarity, and Distractor Plausibility—were rated on a 5-point Likert scale, and Target Population Appropriateness was annotated as a binary label. We report mean Likert scores and population classification accuracy. Inter-rater agreement was measured using Cohen's kappa, along with within-one agreement, defined as the proportion of instances with a rating difference ≤ 1. Further details are provided in Appendix A, Annotation Guidelines.

3.4. Prof-RAG: Profile-aware Retrieval Diagnostic Framework

Alongside PopAnesQA, we introduce Prof-RAG to systematically analyze how query composition influences retrieval behavior. In RAG systems, patient-specific context and clinical question intent are often entangled within a single query, making their effects on retrieval outcomes difficult to isolate. To address this, Prof-RAG decomposes the input query into a question core and a patient profile, enabling controlled query composition and systematic evaluation across varied query configurations. Prof-RAG consists of three stages, as illustrated in Figure 1 (B).

Stage 1. Patient Profile and Question Core Extraction
Stage 1 decomposes the input query into a question core (Q), capturing the primary clinical intent, and a patient profile (P), limited to demographic attributes (e.g., age, sex, and

weight), using an LLM (Prompt A4). Clinical information such as medical history or laboratory findings is explicitly excluded to isolate the influence of demographic context without confounding factors. This retrieval-time patient profile differs from the richer profiles used in benchmark construction, focusing solely on demographic attributes.

Stage 2. Controlled Query Composition for Profile-aware Retrieval
We construct a heuristic retrieval corpus capturing both general and pediatric anesthesiology contexts. We curated approximately 100 domain-relevant keywords to retrieve PubMed abstracts from the past 10 years (Table A3), collecting pediatric-focused documents by augmenting each keyword with pediatric-specific terms ("pediatric", "infant", "children", and "neonate"). After removing overlaps, the final corpus consists of approximately 227K unique abstracts with a near-balanced distribution across general and pediatric populations, ensuring that retrieval trajectory analyses are not confounded by corpus-level population imbalance (see Appendix A). In addition, the source guidelines used for benchmark construction were excluded from the corpus to prevent data leakage. Building on this corpus, we design three query composition strategies to analyze the effect of profile information on retrieval behavior: (1) question core only (Q), (2) combined question core and patient profile (Q+P), and (3) patient profile only (P). Retrieval proceeds through dense retrieval and successive re-ranking, and the three compositions are injected at different steps so that the effect of profile information on each retrieval step can be isolated; the specific staging and retrieval depths are detailed in the Retrieval Pipeline and Ablation Setup in Section 4. For fine-grained analysis, we focus on the top-1 retrieved document after the final stage, allowing us to examine the final evidence selected by the system. This design enables controlled analysis of retrieval behavior across population-specific settings and allows explicit tracking of how population-specific signals influence retrieval outcomes.

Stage 3. Profile-aware Response Generation and Analysis
In the final stage, we perform response generation (Prompts A5–A8) based on the top-1 document retrieved through the profile-aware retrieval pipeline. The selected document is used as supporting evidence to produce answers conditioned on different query compositions. We evaluate the proposed framework across a diverse set of models, including small local models, proprietary API-based models, and medical-specific LLMs, to assess the generality of profile-aware retrieval effects. In addition, we conduct retrieval trajectory analysis by tracking the composition of retrieved documents across different retrieval stages (e.g., dense retrieval and re-ranking). Specifically, we analyze how the proportion of pediatric versus general documents (or both) evolves under different query compositions, providing insights into how profile information influences retrieval behavior.

4. Experimental Setup

Models

We distinguish the models used for benchmark construction, query decomposition, retrieval, and answer generation. The full benchmark construction pipeline (Section 3.2) uses Gemini-3-Flash for all three steps. For query decomposition (Prof-RAG Stage 1, Section 3.4), each evaluated LLM decomposes its own input queries into a patient profile and question

core, rather than relying on a single shared decomposition model. Thus, decomposition is treated as part of the model-specific Prof-RAG pipeline. Retrieval in RAG-based methods uses MedCPT (Jin et al., 2023) as a backbone, with its Query/Article encoders for dense retrieval and Cross-Encoder for re-ranking.

For answer generation, we evaluate a diverse set of LLMs, including general-domain open-source models spanning small to large scales: Mistral-7B-it (Jiang et al., 2023), Llama-3.1-8B-it (Grattafiori et al., 2024), Gemma-3-27B-it (Gemma Team, 2025), Mixtral-8x7B-it (Jiang et al., 2024), and Llama-3.3-70B-it (Grattafiori et al., 2024); medical-domain open-source models SNUH-Hari-q2.5 (SNUH-HARI, 2025) and HuatuoGPT-o1-72B (Chen et al., 2024); and proprietary models GPT-4o (OpenAI, 2024) and Gemini-3-Flash (Google, 2025). For brevity, we omit the "-it" suffix from instruction-tuned model names and refer to all models by their base names unless otherwise specified.

Data

We used PopAnesQA as the primary evaluation dataset, which enables reporting both total accuracy and subgroup accuracy across pediatric, adult, and general subsets. To assess generalizability across different anesthesiology QA distributions and input formats, we additionally evaluated performance on SPA weekly QA (n=219) (Society for Pediatric Anesthesia, 2025) and AnesBench (n=4,343) (Feng et al., 2025). Unlike PopAnesQA, these external datasets do not provide explicit patient profiles, clinical scenarios, and population labels. SPA weekly QA was accessed through membership and used solely for internal evaluation without redistribution. For proprietary models, we report results only on our benchmark due to data access constraints and computational limitations.

Retrieval Pipeline and Ablation Setup

We compare three methods to analyze how profile-aware retrieval affects downstream performance. Baseline denotes direct prompting without retrieval. Base-RAG and Prof-RAG operate over the same 227K retrieval corpus and backbone, differing only in how the query is composed at each step: Base-RAG retrieves and re-ranks with the original, undecomposed query throughout, whereas Prof-RAG performs staged retrieval that progressively incorporates profile information, using Q-only dense retrieval via maximum inner product search (MIPS) to obtain top-64 candidates, cross-encoder re-ranking with $Q{+}P$ to select top-8, and a final P-only re-ranking step to select the top-1 document. This staged design enables systematic analysis of how profile signals influence retrieval behavior at each step.

To isolate the contribution of each component, we conduct ablation studies with several variants. The $Q{\rightarrow}Q$ setting uses only the question core (Q) for both dense retrieval (top-64) and re-ranking (top-1), while the $Q{\rightarrow}Q{+}P$ variant performs re-ranking with combined question and profile but omits the final profile-only re-ranking step. We further evaluate profile robustness by replacing the extracted patient profile with randomly sampled, in-group, or cross-group alternatives. We construct a balanced profile pool by randomly sampling 120 profiles each from pediatric-target and adult-target QAs to ensure balanced representation across populations. This allows us to examine how profile perturbations influence retrieval behavior and downstream performance. Further details are provided in Appendix A.

5. Results

5.1. Benchmark Characteristics and Expert Quality Evaluation

The 623 instances in PopAnesQA cover three population groups (314 pediatric, 159 general, and 150 adult) and five clinical categories, the largest being Pharmacology & Fluid and the smallest Crisis & Complication (Table A1). To mitigate position bias, we balanced the correct-answer options across choices during construction, yielding a near-uniform distribution over A–D. Patient profiles and questions average 144 and 114 characters, respectively.

Human evaluation on 100 sampled instances indicates high overall quality. Average ratings were 4.77±0.62 for guideline consistency, 4.61±0.91 for medical validity, 4.24±1.29 for one-best-answer clarity, and 4.49±0.82 for distractor plausibility, with target population appropriateness reaching 99.5% accuracy (Table A4). Inter-rater agreement was moderate for one-best-answer clarity (Cohen's $\kappa = 0.54$) and accompanied by consistently high within-one agreement (≥ 0.79) across all criteria, indicating that most rating differences were minor. Distractor plausibility showed low κ (0.15) despite near-identical mean ratings across annotators. This reflects a ceiling effect, in which ratings concentrated at 4–5 yield low chance-corrected agreement even when raw agreement is high (the high-agreement/low-κ paradox; Feinstein and Cicchetti, 1990), rather than substantive disagreement, as within-one agreement remained high (0.88).

To complement the Likert-scale ratings, we additionally report a conservative high-confidence summary requiring both annotators to assign a score of at least 4 on each criterion. Under this stricter threshold, 93 instances (out of 100) meet the bar for guideline consistency, 79 for medical validity, 65 for one-best-answer clarity, and 83 for distractor plausibility, with 63 satisfying all four criteria simultaneously. This indicates that guideline consistency is the strongest dimension while one-best-answer clarity is the main quality bottleneck. The close match between the all-criteria rate and the one-best-answer clarity rate suggests that instances meeting the clarity threshold generally also score highly on the others. We release item-level expert ratings so that stricter high-confidence subsets can be constructed as needed.

5.2. LLM Performance and Population Gap

Table 2 summarizes the performance of Baseline, Base-RAG, and Prof-RAG across a diverse set of models. Across all models, we observe a consistent performance disparity across population groups, with pediatric subsets generally exhibiting lower accuracy than adult and general subsets (see Figure A1 for subgroup distributions across model scale). Bootstrap analysis (95% CI, 1,000 resamples) (Efron and Tibshirani, 1994) confirms that this gap remains consistent across bootstrap samples within individual model configurations: for example, Gemma-3-27B under Baseline achieves 63.69 [58.28–69.11] on the pediatric subset versus 77.36 [71.07–83.65] on the general subset, with non-overlapping confidence intervals. A similar trend is observed in proprietary models, where performance on pediatric cases remains markedly lower relative to the general subset, with gaps exceeding 15 percentage points in GPT-4o and approximately 10 percentage points in Gemini-3-Flash. To confirm that this gap is not an artifact of the three-way composition, in which pediatric instances outnumber each of the other groups, we additionally compare pediatric against pooled non-

Table 2: Performance comparison of baseline, Base-RAG, and Prof-RAG across benchmarks. PopAnesQA (n=623) reports subgroup and total accuracies, whereas SPA weekly QA (n=219) and AnesBench (n=4,343) report total accuracy only. Bold and underline indicate the best and second-best within each column among open-source models.

Model	**Method**	**PopAnesQA**				**SPA weekly QA**	**AnesBench**
		Adult	**Pediatric**	**General**	**Total**		
Mistral-7B-it	Baseline	46.67	50.64	61.64	52.49	38.36	46.44
	Base-RAG	52.00	49.68	51.57	50.72	38.36	41.24
	Prof-RAG	50.00	50.32	60.38	52.81	40.64	40.73
Llama-3.1-8B-it	Baseline	60.00	56.05	71.07	60.83	52.51	53.83
	Base-RAG	60.67	54.14	72.33	60.35	43.38	53.42
	Prof-RAG	56.00	56.69	64.78	58.59	47.49	52.94
Gemma-3-27B-it	Baseline	74.00	63.69	77.36	69.66	52.05	60.51
	Base-RAG	73.33	62.42	79.87	69.50	53.88	60.53
	Prof-RAG	<u>74.67</u>	63.06	**85.53**	<u>71.59</u>	52.51	60.28
Mixtral-8x7B-it	Baseline	67.33	59.87	69.81	64.21	48.40	55.45
	Base-RAG	60.67	57.64	59.75	58.91	47.49	49.04
	Prof-RAG	67.33	53.50	67.30	60.35	41.55	48.24
Llama-3.3-70B-it	Baseline	**76.67**	64.33	79.25	71.11	**65.30**	**71.86**
	Base-RAG	74.00	65.29	79.25	70.95	63.93	71.15
	Prof-RAG	72.67	62.42	76.73	68.54	<u>64.84</u>	<u>71.20</u>
Medical-Specific LLMs							
SNUH-Hari-q2.5	Baseline	73.33	<u>67.52</u>	<u>81.76</u>	**72.55**	56.62	64.45
	Base-RAG	68.00	64.65	77.36	68.70	48.86	60.74
	Prof-RAG	68.00	65.92	81.13	70.31	55.25	60.67
HuatuoGPT-o1-72B	Baseline	72.00	63.06	80.50	69.66	61.64	69.10
	Base-RAG	70.00	**67.83**	79.87	71.43	62.10	70.18
	Prof-RAG	69.33	65.29	78.62	69.66	61.64	68.48
Proprietary Models							
GPT-4o	Baseline	76.00	68.15	83.65	74.00	–	–
	Base-RAG	77.33	68.15	81.13	73.68	–	–
	Prof-RAG	77.33	70.70	82.39	75.28	–	–
Gemini-3-Flash	Baseline	90.67	82.48	92.45	87.00	–	–
	Base-RAG	92.00	83.12	90.57	87.16	–	–
	Prof-RAG	92.00	81.85	93.08	87.16	–	–

pediatric (adult+general) accuracy, under which the benchmark is nearly balanced (314 vs. 309). The pediatric gap persists across all methods (e.g., 10.18 points under Baseline, macro-averaged over open-source models; see Table A5), indicating that it reflects a genuine population effect rather than composition imbalance.

Furthermore, Base-RAG often degrades performance relative to Baseline, consistent with prior findings in clinical QA settings. The confidence intervals for Baseline and Base-

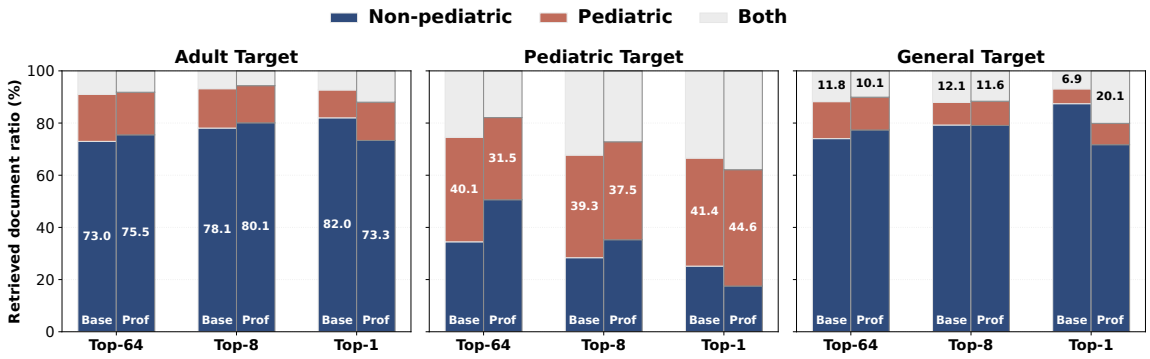

Figure 2: Distribution of retrieved document populations across retrieval stages for Base-RAG and Prof-RAG on Gemma-3-27B. Columns represent target populations (Adult, Pediatric, General). Colors indicate heuristic keyword-based labels: Non-pediatric (general-keyword-only, blue), Pediatric (pediatric-keyword-only, red), and Both (overlap, gray). Only target-aligned proportions are annotated. The trajectory for Llama-3.3-70B is nearly identical and is shown in Figure A2.

RAG largely overlap (e.g., Llama-3.3-70B: 71.11 [67.58–74.48] under Baseline vs. 70.95 [67.58–74.32] with Base-RAG), indicating that these differences are not consistently observed across bootstrap samples. Within RAG-based approaches, Prof-RAG reveals heterogeneous effects compared to Base-RAG, with improvements in some models (e.g., Gemma-3-27B), but limited or even negative effects in others (e.g., Llama-3.3-70B). In particular, Gemma-3-27B achieves the largest gain with Prof-RAG (71.59 total accuracy), performing comparably to 70B-scale models. These heterogeneous effects motivate a deeper analysis of how profile information interacts with retrieval behavior, which we further analyze in the following sections. We note that these RAG results are also sensitive to retrieval configuration: an ablation over retrieval depth (top-1/3/8) and backbone (MedCPT dense/BM25, Robertson and Zaragoza, 2009) shows that several settings surpass the no-retrieval baseline for individual models (e.g., Llama-3.3-70B reaches 72.71 with BM25 top-3, exceeding its Baseline of 71.11), though the best configuration varies by model and backbone (Table A8). Our reported top-1 dense setting is thus a conservative diagnostic choice rather than a per-model optimum. Similar trends hold across additional benchmarks—all models score lower on SPA weekly QA, which is more pediatric-oriented given its pediatric-society origin, than on AnesBench, and RAG-based methods again do not consistently outperform Baseline.

5.3. Retrieval Trajectory Analysis

Figure 2 shows how retrieved document populations evolve across retrieval stages for Base-RAG and Prof-RAG (numerical values in Table A7). We report Gemma-3-27B as a representative model; Llama-3.3-70B, though most degraded by Prof-RAG, follows a nearly identical trajectory (Figure A2), and we note below where the two differ.

Under pediatric targets, Prof-RAG progressively aligns retrieval with the target population, raising pediatric documents from 31.5% (Top-64) to 44.6% (Top-1), whereas Base-RAG stays flat (40.1% to 41.4%). Even at the initial Q-only stage in Prof-RAG, substan-

tially more pediatric documents are retrieved under pediatric than adult targets (31.5% vs. 16.3%), indicating that population signals are partly encoded in the question itself and that the keyword-built corpus retains meaningful population separation.

Under general targets, Prof-RAG instead increases ambiguous "Both"-labeled documents at the final stage (20.1% vs. 6.9% for Base-RAG), suggesting that profile conditioning can blur document selection when the target population is under-specified.

Finally, the two models' trajectories are largely consistent despite independent query decomposition, differing only slightly in intermediate pediatric-target stages. Because they converge at the final stage yet diverge downstream (Gemma improved, Llama degraded), trajectory differences alone cannot explain the performance gap, motivating the qualitative analyses in Section 6.

5.4. Retrieval Pipeline Ablation Results

Figure 3 presents relative performance changes compared to Prof-RAG under retrieval and profile variants (detailed values in Table A6). For Gemma-3-27B, both retrieval variants ($Q \to Q+P$ and $Q \to Q$) result in lower accuracy than Prof-RAG, indicating that the final profile-only re-ranking stage contributes meaningfully to performance gains. Profile variants similarly degrade performance, with cross-group and in-group substitutions falling below Prof-RAG, suggesting that the model is sensitive to the specific profile content used during retrieval. In contrast, Llama-3.3-70B exhibits the opposite pattern: removing profile information progressively improves performance (Prof-RAG $<$ $Q \to Q+P$ $<$ $Q \to Q$), with $Q \to Q$ even surpassing the Baseline. This aligns with the earlier observation that Prof-RAG degrades Llama's performance, and suggests that profile information may act as noise rather than a useful retrieval signal for this model. Among profile variants, random and cross-group substitutions yield similarly low performance as Prof-RAG, while the in-group variant recovers to near-Baseline levels, suggesting that within-population profiles introduce less disruptive noise. Together, these results indicate that the effectiveness of profile-aware retrieval is model-dependent: the same profile information that enhances retrieval alignment in one model can degrade performance in another.

6. Discussion

Our results demonstrate that pediatric performance gaps persist across models and benchmarks, that RAG effects are highly model-specific, and that patient profile information reshapes retrieval patterns in ways that can be both beneficial and detrimental. These findings are consistent with prior work reporting limitations of retrieval-augmented approaches in clinical settings, including the retrieval of task-irrelevant documents through dense retrieval (Dong et al., 2025) and difficulties in handling complex queries that require decomposition (Ammann et al., 2025). Our analysis extends these observations by showing that the entanglement of patient profile information and question intent within a single query may further exacerbate suboptimal retrieval behavior. In particular, our retrieval trajectory and ablation analyses reveal that the same profile information that improves retrieval alignment in Gemma-3-27B acts as disruptive noise in Llama-3.3-70B, indicating that profile-aware retrieval strategies cannot be applied uniformly across models. At the same time, an oracle analysis in which the gold guideline evidence is supplied as the top-1

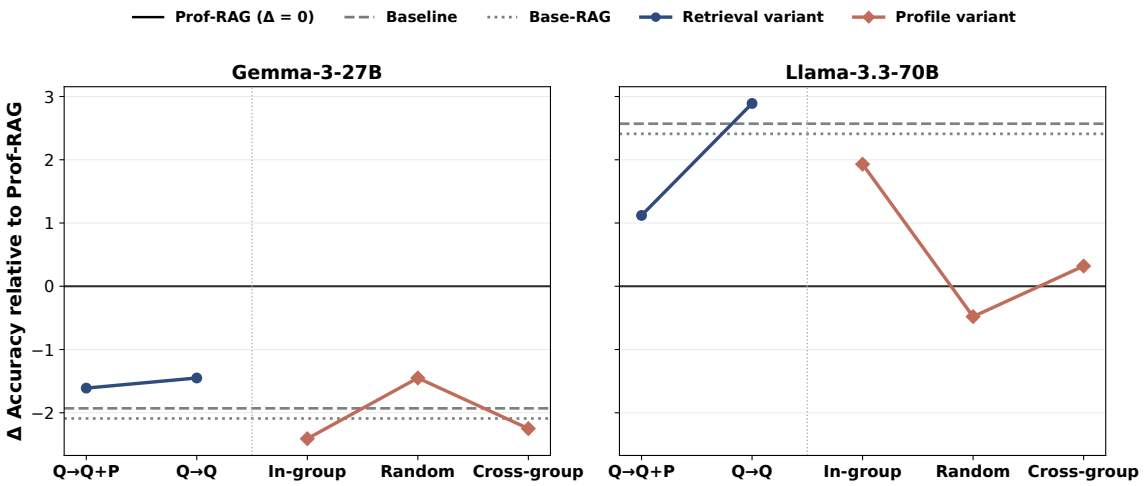

Figure 3: Relative performance (Δ accuracy) compared to Prof-RAG. Retrieval variants (blue) modify the re-ranking stages: $Q{\rightarrow}Q{+}P$ omits the final profile-based stage, and $Q{\rightarrow}Q$ uses only the question core throughout. Profile variants (red) replace the original profile with same-group (In-group), random (Random), or opposite-group (Cross-group) alternatives. Baseline and Base-RAG are indicated by gray dashed lines.

document improves accuracy by 11–19 points and narrows the pediatric gap to under 5, with a slight reversal for SNUH-Hari-q2.5 (Table A9). This evidence is by construction aligned with both the target population and the clinical intent of the question, so the result indicates that the pediatric gap, while reflecting real differences in what models know about pediatric practice, is largely recoverable once such a document reaches the model. What limits current pipelines is thus the ability to select the right evidence rather than the ability to use it. These findings highlight the need for retrieval strategies that balance question intent and patient-specific information, rather than naively combining all query components into a single retrieval input.

Qualitative analysis shown in Figure 4 further demonstrates how retrieval quality directly shapes clinical reasoning and final answer selection. In a pediatric intraoperative fluid management case, Base-RAG retrieved a general review discussing the ongoing debate between isotonic and hypotonic fluids, without providing a clear guideline-level recommendation. As a result, the model relied on incomplete reasoning and incorrectly selected a hypotonic pediatric maintenance fluid, overlooking the risk of perioperative hyponatremia. In contrast, Prof-RAG retrieved a pediatric-specific perioperative guideline explicitly recommending isotonic balanced electrolyte solutions to prevent hyponatremia and maintain physiologic stability. This shifted the model's reasoning from ambiguity to guideline-aligned decision-making, enabling correct selection of isotonic crystalloid fluids. This case highlights that clinically actionable, population-specific evidence—rather than general narrative knowledge—is critical for accurate decision-making, and that profile-aware retrieval can facilitate this transition when properly aligned with clinical intent.

However, the bottom case in Figure 4 reveals how misaligned retrieval can mislead clinical reasoning despite apparent population relevance. In a pediatric case of idiopathic

Success Case: Pediatric Intraoperative Fluid Selection **Prof-RAG: Correct**

Question (abbreviated): A 6-year-old girl (20 kg) undergoing elective orthopedic surgery requires intravenous fluid selection to minimize acute electrolyte complications. **Answer: (D) Isotonic crystalloid solution.**

Base-RAG → Incorrect (C)

Retrieved: "...ongoing debate over administration of isotonic versus hypotonic maintenance fluids..."

Reasoning: Without explicit guidance, chose hypotonic fluid (0.18% NaCl) and overlooked hyponatremia risk.

Prof-RAG → Correct (D)

Retrieved: "...balanced isotonic electrolyte solution with 1–2.5% glucose is recommended... to avoid hyponatremia..."

Reasoning: Explicit isotonic recommendation supported correct selection and flagged the hyponatremia risk of hypotonic options.

Failure Case: Pediatric IPAH with RV Failure **Prof-RAG: Incorrect**

Question (abbreviated): A 7-year-old (22 kg) with IPAH, WHO class IV, RV failure, CI = 1.8 L/min/m^2, and PVR/SVR = 1.1. What is the most appropriate risk assessment and management? **Answer: (C) Higher risk; CCBs contraindicated.**

Base-RAG → Correct (C)

Retrieved: "...assessment of risk and response to therapy is critical in long-term management..."

Reasoning: General PAH review let the model use WHO IV and RV failure to correctly rule out CCBs.

Prof-RAG → Incorrect (B)

Retrieved: "...sildenafil improves preoperative hemodynamics...children with VSD..."

Reasoning: VSD evidence shifted focus to vasodilator testing, overlooking that CCBs are **contraindicated in WHO IV with RV failure**, regardless of vasoreactivity.

Figure 4: Qualitative retrieval analysis. **Top:** Profile-aware retrieval recovers a population-specific perioperative guideline, enabling correct fluid management. **Bottom:** Profile signals retrieve a population-matched but disease-misaligned document (VSD vs. IPAH), so population alignment without intent alignment yields a contraindicated recommendation.

pulmonary arterial hypertension (IPAH) with severe right ventricular failure, Base-RAG retrieved a general pulmonary arterial hypertension (PAH) review that preserved the core clinical reasoning pathway, allowing the model to correctly identify the patient as high-risk and recognize calcium channel blockers as contraindicated. In contrast, Prof-RAG prioritized pediatric-specific evidence but retrieved a study on ventricular septal defect (VSD)–related pulmonary hypertension, which is clinically distinct from IPAH. This misalignment shifted the model's reasoning toward vasodilator responsiveness and preoperative hemodynamic optimization, leading it to recommend calcium channel blockers — a clinically contraindicated action in this setting. Critically, this error arose not from a lack of population awareness, but from overemphasis on demographic similarity at the expense of disease-specific clinical intent. This case highlights that retrieval driven primarily by patient profile can introduce clinically unsafe reasoning paths, underscoring the need to balance population alignment with task-specific relevance in clinical decision-making. These cases demonstrate that retrieval quality, rather than model reasoning alone, can critically influence clinical decision outcomes (see Appendix Retrieval Case Analysis for additional cases).

7. Conclusion

In this work, we present PopAnesQA, a population-aware benchmark, and Prof-RAG, a diagnostic framework for examining how patient profile information shapes retrieval in clinical decision-making. Our results reveal a consistent pediatric performance gap and highlight the unstable, model-dependent behavior of standard RAG in population-sensitive settings. We further show that incorporating patient profiles introduces a fundamental trade-off: it can improve retrieval of population-relevant evidence, but may also bias the model toward clinically misaligned information. These findings emphasize that effective clinical retrieval requires jointly accounting for question intent and patient-specific context, and that population-aware evaluation is essential to identify failure modes before clinical deployment.

Limitations and Future Work

This study has several limitations. First, our benchmark focuses on pediatric anesthesiology, a relatively specialized clinical setting, and its items are formatted as multiple-choice questions, which abstract away the dynamic nature of real clinical decision-making. Nevertheless, because each item embeds a patient profile and targets an action-oriented management decision rather than isolated knowledge recall, PopAnesQA stays closer to real-world decision-making than standard knowledge-based QA. Second, the benchmark was constructed using an LLM-based pipeline, raising circularity concerns when evaluating other LLMs. To mitigate this, the pipeline grounds all generated content in source guideline text rather than relying on the model's parametric knowledge, and most evaluated models differ from the generation model. Expert evaluation on 100 sampled instances further confirmed high guideline consistency and medical validity, though it may not fully capture variability across the dataset. Finally, the retrieval corpus was labeled using keyword-based heuristics, so its population labels should be treated as approximate proxies.

These limitations point to several directions for future work. While the framework is demonstrated on pediatric anesthesiology, extending and empirically validating it across other population-sensitive axes (e.g., geriatric, obstetric) and specialties is an important next step. Because each benchmark item derives from a known source guideline, this evidence can serve as gold retrieval target, enabling a finer-grained trajectory analysis of whether Prof-RAG progressively ranks the correct guideline evidence higher across retrieval stages. In addition, manually validating a sample of the corpus population labels would quantify the noise introduced by keyword-based construction. Finally, retrievers explicitly optimized for patient-context alignment, together with prompt-perturbation analyses, would clarify how stable the observed population gaps and retrieval effects are.

Acknowledgments

This research was supported by grant No. KSPA-2024-001 from the Korean Society of Pediatric Anesthesiologists. This research was also supported by the NAVER Digital Bio Innovation Research Fund, funded by NAVER Corporation (Grant No. 3720252600), and supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2025-02213749).

References

Paul J. L. Ammann, Jonas Golde, and Alan Akbik. Question decomposition for retrieval-augmented generation. In Jin Zhao, Mingyang Wang, and Zhu Liu, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 497–507, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-254-1. doi: 10.18653/v1/2025.acl-srw.32. URL https://aclanthology.org/2025.acl-srw.32/.

Rajesh Bhayana. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*, 310(1):e232756, 2024.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, 2025.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Jurgen C De Graaff, Mathias Fuglsang Johansen, Martinus Hensgens, and Thomas Engelhardt. Best practice & research clinical anesthesiology: safety and quality in perioperative anesthesia care. update on safety in pediatric anesthesia. *Best Practice & Research Clinical Anaesthesiology*, 35(1):27–39, 2021.

Gregory R Dobson, Anthony Chau, Justine Denomme, Samantha Frost, Giuseppe Fuda, Conor Mc Donnell, Robert Milkovich, Andrew D Milne, Kathryn Sparrow, Yamini Subramani, et al. Guidelines to the practice of anesthesia—revised edition 2025. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie*, 72(1):15–63, 2025.

Xuanzhao Dong, Wenhui Zhu, Hao Wang, Xiwen Chen, Peijie Qiu, Rui Yin, Yi Su, and Yalin Wang. Talk before you retrieve: Agent-led discussions for better rag in medical qa. *arXiv preprint arXiv:2504.21252*, 2025.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Alvan R Feinstein and Domenic V Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549, 1990.

Xiang Feng, Wentao Jiang, Zengmao Wang, Yong Luo, Pingbo Xu, Baosheng Yu, Hua Jin, Bo Du, and Jing Zhang. Anessuite: A comprehensive benchmark and dataset suite for anesthesiology reasoning in llms. *arXiv preprint arXiv:2504.02404*, 2025.

Chase M Fensore, Rodrigo M Carrillo-Larco, Megha K Shah, and Joyce C Ho. Does domain-specific retrieval augmented generation help llms answer consumer health questions? In *Machine Learning for Healthcare Conference*. PMLR, 2025.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

David M Gaba. Dynamic decision-making in anesthesiology: cognitive models and training approaches. In *Advanced models of cognition for medical training and practice*, pages 123–147. Springer, 1992.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Google. Gemini 3 flash: Frontier intelligence built for speed, 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/. Accessed: 2026-04-07.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Jason Holmes, Zhengliang Liu, Lian Zhang, Yuzhen Ding, Terence T Sio, Lisa A McGee, Jonathan B Ashman, Xiang Li, Tianming Liu, Jiajian Shen, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Frontiers in Oncology*, 13:1219326, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Hyunjae Kim, Jiwoong Sohn, Aidan Gilson, Nicholas Cochran-Caggiano, Serina Applebaum, Heeju Jin, Seihee Park, Yujin Park, Jiyeong Park, Seoyoung Choi, et al. Rethinking retrieval-augmented generation for medicine: A large-scale, systematic expert evaluation and practical insights. *arXiv preprint arXiv:2511.06738*, 2025.

Siun Kim and Hyung-Jin Yoon. Questioning our questions: How well do medical qa benchmarks evaluate clinical capabilities of language models? In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 274–296, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Tingjun Liu, Xucheng Wang, Matthew Inkman, Julian C Hong, Michael R Waters, and Jin Zhang. Development of a rag-based expert llm for clinical support in radiation oncology. *medRxiv*, pages 2025–09, 2025.

Ming-Jie Luo, Jianyu Pang, Shaowei Bi, Yunxi Lai, Jiaman Zhao, Yuanrui Shang, Tingxin Cui, Yahan Yang, Zhenzhe Lin, Lanqin Zhao, et al. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. *JAMA ophthalmology*, 142(9):798–805, 2024.

Fnu Neha, Deepshikha Bhati, and Deepak Kumar Shukla. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. *AI*, 6(9):226, 2025.

OpenAI. Gpt-4o system card, 2024.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

Mahimai Raja, E Yuvaraajan, et al. A rag-based medical assistant especially for infectious diseases. In *2024 International Conference on Inventive Computation Technologies (ICICT)*, pages 1128–1133. IEEE, 2024.

Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc, 2009.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

SNUH-HARI. hari-q2.5, May 2025. URL https://huggingface.co/snuh/hari-q2.5. Healthcare AI Research Institute(HARI) of Seoul National University Hospital(SNUH).

Society for Pediatric Anesthesia. Questions of the week. https://pedsanesthesia.org/questions-of-the-week/, 2025. Accessed through SPA membership. Used solely for internal evaluation.

Marjorie Podraza Stiegler and Avery Tung. Cognitive processes in anesthesiology decision making. *Anesthesiology*, 120(1):204–217, 2014.

Michael RJ Sury, Renuka Arumainathan, Alla M Belhaj, James H MacG Palmer, Tim M Cook, and Jaideep J Pandit. The state of uk pediatric anesthesia: a survey of national health service activity. *Pediatric Anesthesia*, 25(11):1085–1092, 2015.

Hang Sheung Wong and Tsz Kwan Wong. Multi-evidence clinical reasoning with retrieval-augmented generation for emergency triage: Retrospective evaluation study. *JMIR Medical Informatics*, 14(1):e82026, 2026.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, 2024a.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific, 2024b.

Huan Zhao, Qian Ling, Yi Pan, Tianyang Zhong, Jin-Yu Hu, Junjie Yao, Fengqian Xiao, Zhenxiang Xiao, Yutong Zhang, San-Hua Xu, et al. Ophtha-llama2: A large language model for ophthalmology. *arXiv preprint arXiv:2312.04906*, 2023.

Yuxuan Zhou, Xien Liu, Chenwei Yan, Chen Ning, Xiao Zhang, Boxun Li, Xiangling Fu, Shijin Wang, Guoping Hu, Yu Wang, et al. Evaluating llms across multi-cognitive levels: From medical knowledge mastery to scenario-based problem solving. *arXiv preprint arXiv:2506.08349*, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

Appendix A.

Implementation Details

The LLM used for benchmark construction was selected after comparing two proprietary models, GPT-4o and Gemini-3-Flash. An anesthesiology specialist blindly reviewed approximately 20 generated samples from each model, and Gemini-3-Flash was selected for the final pipeline. All experiments were conducted using `Python` (v3.9.19) and `PyTorch` (v2.8.0). For open-source models, inference was performed using Hugging Face `transformers` (v4.57.1) and accelerated with `vLLM` (v0.10.2). To ensure consistency across experiments, all generation settings were fixed, with temperature set to 0, prefix caching disabled, a fixed random seed, and a batch size of 24. The maximum generation length was set to 3,000 tokens for final answer generation and 512 tokens for profile decomposition, and was kept consistent across models. All experiments were conducted on up to four NVIDIA A6000 GPUs.

Query composition was performed by separating the input into a question core (Q) and a patient profile (P), and for the combined condition (Q+P), concatenating them in the form "P Q" for retrieval and re-ranking. Dense retrieval uses the MedCPT Query and Article encoders with CLS pooling (FP32) via MIPS, followed by MedCPT Cross-Encoder re-ranking. Abstracts are chunked to a maximum of 384 tokens with 64-token overlap (272 tokens on average); since the encoders accept up to 512 tokens, capping chunks at 384 allows a query (mean length 101 tokens) and a chunk to jointly fit within the 512-token limit during re-ranking, reducing truncation.

Retrieval Corpus Construction

We constructed the retrieval corpus by collecting PubMed abstracts using domain-relevant keywords. Each keyword was submitted as an independent PubMed query to collect up to 1,500 abstracts per keyword, with a date filter from January 1, 2015 to December 31, 2024. Queries were executed in batches (size=250) with pagination (retstart) to ensure consistent coverage. General and pediatric-augmented keyword sets yielded 115,680 and 129,473 abstracts, respectively, with 18,369 overlapping between the two sets. After deduplication, the final corpus consists of approximately 227K unique abstracts. This near-balanced collection across general and pediatric queries ensures that observed retrieval patterns are not driven by corpus-level population imbalance.

To ensure a rigorous evaluation of models' retrieval and reasoning capabilities, we explicitly excluded the source guidelines used for benchmark construction from the corpus. Initial exploratory experiments showed that including these source documents led to inflated performance due to data leakage. By removing them, the reported results reflect the models' ability to synthesize information from related abstract-level clinical evidence rather than retrieving the exact source material.

Supplementary Figures and Tables

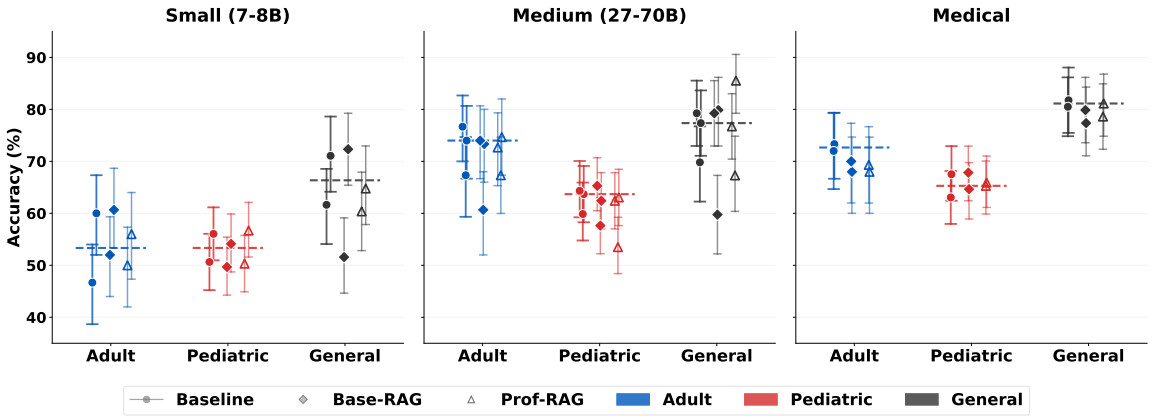

Figure A1: Distribution of subgroup accuracy across population groups (Adult, Pediatric, General), stratified by model scale: Small (Mistral-7B, Llama-3.1-8B), Medium (Gemma-3-27B, Mixtral-8x7B, Llama-3.3-70B), and Medical (SNUH-Hari, HuatuoGPT-o1). Each point represents a model under a given method: Baseline (circle), Base-RAG (diamond), and Prof-RAG (triangle, open), all shown with 95% bootstrap confidence intervals. Dashed lines indicate the baseline performance (median across models) within each group. Across all model scales, pediatric accuracy is consistently lower than adult and general accuracy, confirming that the population gap persists regardless of model size or domain specialization.

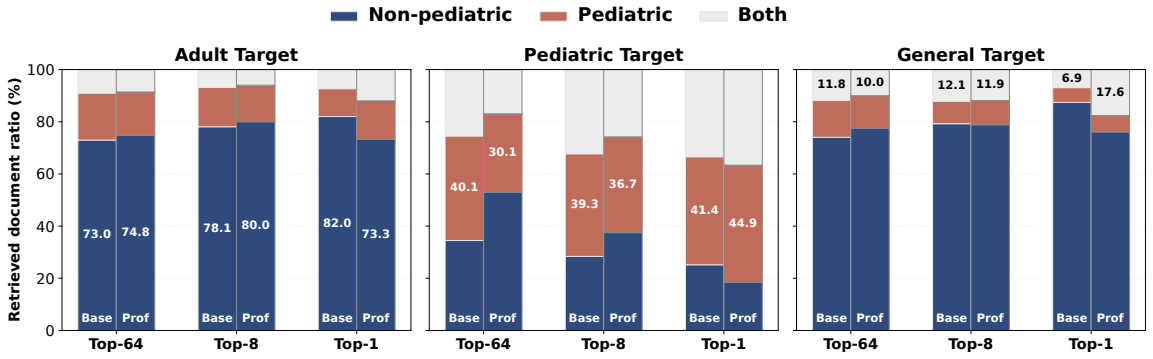

Figure A2: Distribution of retrieved document populations across retrieval stages for Base-RAG and Prof-RAG on Llama-3.3-70B. Columns represent target populations (Adult, Pediatric, General). Colors indicate Non-pediatric (blue), Pediatric (red), and Both (gray) document proportions. Only target-aligned proportions are annotated. The trajectory closely mirrors that of Gemma-3-27B (Figure 2), despite each model performing its own query decomposition.

Table A1: Statistics of the proposed PopAnesQA (623 instances). Correct-answer options were balanced across A–D during construction to mitigate position bias. String lengths are measured in characters.

Category	**Subgroup**	**Count**
Target Population	Pediatric	314
	General	159
	Adult	150
Clinical Category	Pharmacology & Fluid	185
	Airway & Equipment	175
	Pre-op Assessment	116
	Post-op & Pain	91
	Crisis & Complication	56
Correct Answer	A	156
	B	156
	C	156
	D	155
Field		**Mean ± Std (chars)**
Patient Profile		144 ± 67
Question		114 ± 28
Clinical Scenario		263 ± 89
Explanation		821 ± 209

Table A2: List of guideline documents used in benchmark construction, identified by PubMed IDs (PMIDs), DOIs, or titles.

PMID / DOI / Title
27290980; 35065393; 36427486; 38325249; 38843649; 38992466; 38124208; 34762729; 37533337; 38018248; 38699880; 31792941; 30829672; https://doi.org/10.17085/apm.2018.13.1.18; 33329766; 34352966; 36317427; 37919917; 37205803; 37061429; 26556848; 31729018; 34251028; 32467512; 32017012; 30378102; 33413977; 24951446; 31096732; 31163107; 31591858; 32209961; 37073521; 37750295; 38228393; https://doi.org/10.4097/kja.d.18.00073.1; 30086609; 29969889; 30513567; Paediatric Anaesthesia For Beginners; 34397527; 37417808; 28177176; 31435997; 31538393; 31437328; 32298502; 34403548; 34837438; 34741785; 34758163; 34878697; 34902201; 34877746; 34877744; 34379843; 35293066; 36178188; 36017580; 36424875; 36264219; 35266610; 35656935; 35993398; 35150181; 36876548; 37326251; 38146668; 38146636; 37650686; 37800662; 37199294; 37449338; 38462998; 39148245; 39403896; 38546348; 38980227; 26534956; 37972480; 39043129; 33812665; 37720558; 38481418; 38899313; 39099754; 37588272; 29334501; 28045707; 33201518; 34143444; 30161070; 30929307;

Table A3: Anesthesia-related keyword dictionary grouped by category. Each entry is represented as a semicolon-separated keyword string.

Category	**Keyword string**
Airway	airway management; difficult intubation; difficult airway; intubation; endotracheal tube; perioperative respiratory adverse event; hypoxemia; hypoxia; videolaryngoscope; cricoid pressure; supraglottic airway devices; laryngeal mask airway; laryngospasm; high flow nasal oxygen; extubation; mask ventilation; facemask ventilation; airway obstruction; rapid sequence induction; glidescope; fiberoptic intubation; front-of-neck access; ventilation; one lung ventilation; lung isolation
Drugs	volatile anesthesia; total intravenous anesthesia; neuromuscular block; alfentanil; remifentanil; sufentanil; fentanyl; dexmedetomidine; propofol; ketamine; thiopental; midazolam; rocuronium; succinylcholine; cisatracurium; vecuronium; sevoflurane; enflurane; isoflurane; nitrous oxide; desflurane; analgesia; sugammadex; target controlled infusion; opioid; muscle relaxant; reversal agent
Regional	nerve block; plane block; plexus block; caudal anesthesia
General	general anesthesia; monitored anesthesia care; regional anesthesia; spinal anesthesia; epidural anesthesia; procedural sedation; nonoperating room anesthesia; surgery; anesthesia; induction; maintenance; emergence; anesthetic; undergoing surgery; operating room
Ventilator	anesthetic machine; breathing circuit
NPO/Preoperative	gastric emptying time; preoperative fasting; premedication; preoperative; gastric ultrasound; pre-anesthesia
Perioperative	perioperative; intraoperative; fluid responsiveness
Postoperative	emergence delirium; postoperative pain; emergence agitation; postoperative nausea and vomiting
Miscellaneous	anesthetic neurotoxicity; enhanced recovery after surgery; cardiopulmonary resuscitation
Intraoperative	cardiac arrest; resuscitation; central venous catheterization; arterial catheterization; arterial cannulation; central venous cannulation; central catheter; vascular access

Table A4: Human evaluation results on 100 sampled instances from PopAnesQA. Average rating is reported as mean ± standard deviation. Inter-rater agreement is measured using quadratic weighted Cohen's kappa for Likert-scale metrics (5-point Likert scale). Within-one agreement indicates the proportion of ratings differing by at most one point. Target population appropriateness is binary (Correct/Incorrect) and reported as accuracy.

Metric	**Average Rating**	**Inter-rater Agreement**	**Within-one Agreement**
Guideline Consistency	4.77 ± 0.62	0.56	0.97
Medical Validity	4.61 ± 0.91	0.41	0.85
One-Best-Answer Clarity	4.24 ± 1.29	0.54	0.79
Distractor Plausibility	4.49 ± 0.82	0.15	0.88
Target Population Appropriateness	99.50	–	–

Table A5: Pediatric vs. non-pediatric (adult+general) accuracy on PopAnesQA, macro-averaged over the open-source models. The pediatric gap persists under this near-balanced grouping (314 vs. 309 instances).

Method	**Pediatric**	**Non-pediatric**	**Gap**
Baseline	60.74	70.92	10.18
Base-RAG	60.24	68.56	8.32
Prof-RAG	59.60	69.58	9.98

Table A6: Ablation study on profile-aware retrieval strategies on PopAnesQA. This table reports the overall performance corresponding to Figure 3. The best and second-best performances within each model and column are highlighted in bold and underlined, respectively.

Model	**Method**	**PopAnesQA**			
		Adult	**Pediatric**	**General**	**Total**
Gemma-3-27B-it	Baseline	74.00	<u>63.69</u>	77.36	69.66
	Base-RAG	73.33	62.42	79.87	69.50
	Prof-RAG	<u>74.67</u>	63.06	**85.53**	**71.59**
	Q→Q+P	73.33	63.06	80.50	69.98
	Q→Q	74.00	63.38	79.87	<u>70.14</u>
	Random Profile	<u>74.67</u>	62.42	<u>81.13</u>	<u>70.14</u>
	Cross-group	**76.00**	62.74	76.10	69.34
	In-group	70.67	**64.01**	77.99	69.18
Llama-3.3-70B-it	Baseline	<u>76.67</u>	64.33	79.25	<u>71.11</u>
	Base-RAG	74.00	<u>65.29</u>	79.25	70.95
	Prof-RAG	72.67	62.42	76.73	68.54
	Q→Q+P	72.67	64.65	76.73	69.66
	Q→Q	73.33	**66.56**	79.25	**71.43**
	Random Profile	72.00	59.87	**80.50**	68.06
	Cross-group	71.33	62.10	<u>79.87</u>	68.86
	In-group	**77.33**	62.74	79.25	70.47

Table A7: Retrieval trajectory analysis of Base-RAG and Prof-RAG across different target populations. For each target group, we report the percentage of retrieved documents belonging to the non-pediatric, pediatric, and overlapping (both) corpora at different retrieval stages (Top-64, Top-8, and Top-1).

Model	**Target**	**Stage**	**Method**	**Retrieved document distribution (%)**		
				Non-pediatric	**Pediatric**	**Both**
Gemma-3-27B-it	Adult	Top-64	Base-RAG	72.98	18.01	9.01
			Prof-RAG	75.47	16.34	8.19
		Top-8	Base-RAG	78.08	15.08	6.83
			Prof-RAG	80.08	14.25	5.67
		Top-1	Base-RAG	82.00	10.67	7.33
			Prof-RAG	73.33	14.67	12.00
	Pediatric	Top-64	Base-RAG	34.47	40.06	25.46
			Prof-RAG	50.60	31.50	17.89
		Top-8	Base-RAG	28.38	39.33	32.29
			Prof-RAG	35.27	37.54	27.19
		Top-1	Base-RAG	25.16	41.40	33.44
			Prof-RAG	17.52	44.59	37.90
	General	Top-64	Base-RAG	74.03	14.17	11.80
			Prof-RAG	77.35	12.57	10.08
		Top-8	Base-RAG	79.25	8.65	12.11
			Prof-RAG	79.09	9.28	11.64
		Top-1	Base-RAG	87.42	5.67	6.92
			Prof-RAG	71.70	8.18	20.13
Llama-3.3-70B-it	Adult	Top-64	Base-RAG	72.98	18.01	9.01
			Prof-RAG	74.80	16.60	8.59
		Top-8	Base-RAG	78.08	15.08	6.83
			Prof-RAG	80.00	14.00	6.00
		Top-1	Base-RAG	82.00	10.67	7.33
			Prof-RAG	73.33	14.67	12.00
	Pediatric	Top-64	Base-RAG	34.47	40.06	25.46
			Prof-RAG	52.90	30.13	16.94
		Top-8	Base-RAG	28.38	39.33	32.29
			Prof-RAG	37.50	36.70	25.80
		Top-1	Base-RAG	25.16	41.40	33.44
			Prof-RAG	18.47	44.90	36.62
	General	Top-64	Base-RAG	74.03	14.17	11.80
			Prof-RAG	77.52	12.46	10.02
		Top-8	Base-RAG	79.25	8.65	12.11
			Prof-RAG	78.85	9.28	11.87
		Top-1	Base-RAG	87.42	5.67	6.92
			Prof-RAG	76.10	6.29	17.61

Table A8: Retrieval depth (top-1/3/8) and backbone (MedCPT Dense/BM25) ablation for Gemma-3-27B and Llama-3.3-70B on PopAnesQA. Total and subgroup accuracies are reported; the no-retrieval baseline is shown for reference. Within each model, bold and underline indicate the best and second-best value in each column. Several configurations exceed the no-retrieval baseline, though the best setting varies by model, depth, and backbone.

Model	**Method**	**Retrieval**	**Top-k**	**Total**	**Adult**	**Pediatric**	**General**
Gemma-3-27B	Baseline	–	–	69.66	<u>74.00</u>	63.69	77.36
	Base-RAG	Dense	1	69.50	73.33	62.42	79.87
			3	69.02	**74.67**	62.74	76.10
			8	68.06	71.33	63.38	76.10
		BM25	1	<u>70.95</u>	73.33	64.33	<u>81.76</u>
			3	69.66	<u>74.00</u>	63.69	77.36
			8	66.77	69.33	62.10	73.58
	Prof-RAG	Dense	1	**71.59**	**74.67**	63.06	**85.53**
			3	70.63	73.33	<u>64.65</u>	79.87
			8	69.98	70.00	**66.88**	74.21
		BM25	1	68.38	**74.67**	60.51	77.99
			3	69.66	73.33	64.01	77.36
			8	67.74	68.67	62.10	77.99
Llama-3.3-70B	Baseline	–	–	71.11	**76.67**	64.33	79.25
	Base-RAG	Dense	1	70.95	74.00	65.29	79.25
			3	71.43	<u>76.00</u>	64.01	79.25
			8	70.14	74.00	64.01	78.62
		BM25	1	69.34	72.00	63.06	79.25
			3	<u>72.39</u>	74.00	<u>65.92</u>	**83.65**
			8	64.69	68.00	59.55	71.70
	Prof-RAG	Dense	1	68.54	72.67	62.42	76.73
			3	70.31	72.00	65.29	78.62
			8	70.63	73.33	64.01	81.13
		BM25	1	70.63	<u>76.00</u>	64.65	77.36
			3	**72.71**	**76.67**	**66.24**	<u>81.76</u>
			8	65.01	70.00	59.24	71.70

Table A9: Oracle gold-evidence analysis on the PopAnesQA, evaluated on four representative models spanning open-source, medical, and proprietary categories. Providing the gold guideline evidence as the top-1 retrieved document (an upper-bound setting) substantially improves accuracy over baseline across representative models, indicating that the bottleneck lies in retrieving appropriate evidence rather than in the models' inability to use external knowledge. Subgroup accuracies are for the gold-evidence setting.

Model	**Baseline**	**Prof-RAG**	**GT-evidence**	**GT-evidence subgroup**			**Δ vs. Baseline**
			top-1	**Adult**	**Pediatric**	**General**	
Gemma-3-27B	69.66	71.59	89.09	92.00	86.62	91.20	+19.43
Llama-3.3-70B	71.11	68.54	82.34	84.67	80.89	83.02	+11.23
SNUH-Hari-q2.5	72.55	70.31	88.60	85.33	89.17	90.57	+16.05
GPT-4o	74.00	75.28	91.17	92.00	90.45	91.82	+17.17

Prompts

Prompt A1. Step 1 – Guideline Content Classification

[System Prompt]

```
You are a clinical text reviewer specializing in Anesthesiology.
```

[User Prompt]

```
### Instruction:
Analyze the following OCR-extracted guideline text.  Determine whether it contains
practical clinical directives or recommendations relevant to Anesthesiology
for either Pediatric or Adult patients.
A text is considered RELEVANT only if it includes at least one of the following:
    • perioperative assessment or optimization
    • airway evaluation or airway device selection/management
    • anesthetic or sedative drug use, dosing, adjustment, or monitoring
    • intraoperative or postoperative management related to anesthesia care
    • fluid, blood product, or anticoagulant management in the perioperative setting
    • pain management where anesthesiology is responsible
    • anesthetic emergency or complication recognition/management

The text is NOT relevant if it is:
    • purely administrative (billing, scheduling, consent wording, references, citations)
    • general pathophysiology without anesthetic management implications
    • nursing workflow not specific to anesthesia practice
    • purely equipment vendor description
    • background educational material without clinical action

If relevant, classify the text into ONE best-fit category:
Categories
    • Pre-op Assessment:  NPO guidelines, anxiety, premedication, cardiac/pulmonary risk
      evaluation
    • Pharmacology & Fluid:  Drug dosing, fluid therapy, transfusion, anticoagulants/DOACs
    • Airway & Equipment:  Tube size/depth, airway approach, difficult airway plans, ventilator
      settings
    • Crisis & Complication:  Laryngospasm, bronchospasm, LAST, malignant hyperthermia, arrest,
      hypotension
    • Post-op & Pain:  Pain control, PONV, emergence delirium/agitation, PACU discharge

Target Population Classification
Classify the target population as one of:  "Pediatric", "Adult", or "General".
Rules:
    • Pediatric:  explicit children OR weight/age-based practice
    • Adult:  adult-specific disease/context OR unsafe for pediatric use
    • General:  applies safely to both without modification

When in doubt:
    • If applicable to both → General
    • Otherwise → Adult

Ignore OCR noise.
Output JSON format
{EXPECTED JSON FORMAT}
Text to analyze:
{input_text}
```

Prompt A2. Step 2 – Clinical Logic Extraction

[System Prompt]

You are a Medical Data Auditor specializing in Anesthesiology. Your role is to extract ONLY explicit, verifiable clinical logic from text. Do NOT infer or invent missing information.

[User Prompt]

Instruction:
Based on the previously-assigned classification:
- target_population = "{*target_population*}"
- category = "{*category*}"

Extract ONLY clinical logic that is explicit, actionable, and rule-based. A statement qualifies as a clinical rule ONLY if it clearly changes how a clinician should prescribe, dose, time, monitor, or perform a clinical procedure.

VERY IMPORTANT STRICT RULES

1. NO EXTERNAL KNOWLEDGE OR INFERENCE

- Extract ONLY what is explicitly stated in the text. Statements such as "should", "must", "do not", "is recommended" count ONLY if they clearly change management.
- If no clear clinical instruction exists, return an empty rules list.
- Do NOT extract caution-only statements (e.g., "use caution", "monitor closely") unless a concrete management action is specified.

2. ALLOWED CONTENT

A statement qualifies as a rule if it clearly defines a clinical decision or instruction.
Typical examples include:

- when to start, stop, hold, or restart a drug
- drug dose or adjustment rules
- timing or dose changes based on conditions
- measurable thresholds triggering actions
- procedural or crisis response steps
- airway or device selection/settings
- postoperative treatment or monitoring instructions

3. EXCLUDED CONTENT

Do NOT extract:

- descriptive background information
- statistical associations
- mechanistic or pharmacokinetic explanations
- general risk or safety statements
- statements without explicit clinical action
- ambiguous or incomplete numeric content

4. CATEGORY-BOUND EXTRACTION

- Extract ONLY rules relevant to the assigned category and population.
- If spanning multiple categories, include ONLY if the primary action matches the assigned category.

5. DO NOT MERGE DIFFERENT VARIABLES

Treat timing, restart intervals, half-life, washout, maintenance dose, and bolus dose as separate concepts.

6. RULE GROUPING

If multiple conditions lead to the SAME action, represent them as ONE rule with multiple triggers.

7. NO APPROXIMATION

- Preserve exact numbers and inequality symbols; Do NOT convert units.

8. OCR NOISE RULE

Exclude ambiguous or corrupted values.

9. CONDITION_BASE ANCHORING RULE

- Anchor to the exact drug, device, procedure, or population.; Use patient-centered phrasing when implied (e.g., "patients receiving X").; Do NOT generalize or invent broader contexts. ;Match the narrowest explicit scope.

Output JSON format
{*EXPECTED JSON FORMAT*}
Do NOT include invented rules.
Do NOT return commentary outside the JSON.

Now extract rules strictly for {*target_population*} in this category: {*category*}
SOURCE TEXT:
{*input_text*}

Prompt A3. Step 3 – Question-Answer Pair Generation

[System Prompt]

You are a Board Exam Question Creator specializing in Anesthesiology. Your job is to:
Create realistic, high-quality multiple-choice clinical vignettes. Strictly follow the provided guideline logic. Never invent medical rules, numbers, or thresholds. Never contradict the provided source text. Prioritize clinical realism and educational value. You must:
Treat all provided rules as authoritative. Use only the given rules and source text. Avoid adding outside knowledge. Ensure the correct answer is unambiguously supported by the rules and source text.
Multi-rule policy (general, non-case-specific):
If multiple rules are provided, combine them into ONE MCQ only when they can naturally occur in a single realistic scenario WITHOUT adding assumptions. Do NOT force unrealistic co-occurrence.
If rules seem incompatible, mutually exclusive, or would require invented conditions, do NOT merge; instead build a single-rule MCQ using the most clearly supported rule. When merging is feasible, design the item to require rule-interaction reasoning (prioritization, exceptions, conflicts, completeness), not mere keyword matching. Strict grounding: never invent facts/thresholds/management steps beyond the provided rules/source text.

[User Prompt]

Task Context: You must create an anesthesiology board-style MCQ that applies ONLY to:
- Target population = "{*target_population*}"
- Topic category = "{*category*}"
Do NOT use rules from other populations or categories.

Guideline Logic (Authoritative): This is the extracted clinical logic you MUST follow:
{*rules_json*}
Each rule contains:
The payload contains either a single rule or a list of rules:

- If "rules" is a list, each rule contains:
- condition_base: whowhat the rule applies to
- action: what must be done
- triggers: a list of one or more triggering conditions
- source_fragment: exact supporting text
- You may also see, merge_preference: guidance on whether to merge rules, merge_goal: what "good merging" means

You may ONLY use these rules to build the case. --- ### Source Text (For Validation) Use this to verify wording and numbers: {*input_text*}
If there is any mismatch between rules and text, always trust the SOURCE TEXT.

Instruction:
You are a Board Exam Question Creator for {*target_population*} Anesthesiology.
Based on the provided Guideline Logic and Source Text, create ONE high-quality MCQ.
Requirements:
0. Multi-rule decision (only if multiple rules are provided)

- FIRST decide if you can realistically combine two or more rules into a single vignette using ONLY the given triggers/conditions and source text.
- If YES: create a System-2-like MCQ that requires applying at least TWO rules together via rule-interaction reasoning:
  - prioritization (which action matters most when multiple triggers apply)
  - exception handling (when one rule modifies/limits another)
  - conflict resolution (avoid choices that satisfy one rule but violate another)
  - completeness (best option addresses all applicable triggers; distractors miss one key trigger)
- If NO: do NOT force merging. Create a single-rule MCQ using the rule with the clearest triggers and strongest direct support in the source text.

1. Patient Profile Create a realistic {*target_population*} patient.

- If Pediatric: specify age (e.g., 4yo), weight (e.g., 16kg).
- If Adult: specify age (e.g., 65yo), comorbidities, weight if relevant.

Prompt A3. Continued.

2. Scenario
   - Must strictly follow the extracted rules.
   - Must clearly activate at least one rule via its triggers.
   - Do not add extra conditions not present in the rules.
   - If merging rules: ensure the scenario naturally activates multiple rules without introducing assumptions beyond the given rules/source text.

3. Distractors
   - Create three plausible but incorrect options.
   - Try to vary error types when possible:
     - Wrong time, dose, or threshold
     - Applying a rule from a different population or context
     - Outdated or superseded practice
     - Overly aggressive or unsafe action
     - Ignoring a key condition or trigger
   - Do NOT force rigid categories.
   - Choose the most realistic and educational errors.
   - When natural:
     - Include at least one common clinical mistake.
     - Include at least one unsafe or contraindicated option.
   - If artificial, prioritize realism.

Wording constraint (MUST):
   - Do NOT use phrases like "According to the guideline(s)", "per the guideline", "based on the guideline", "current guidelines recommend", or similar wording in the question stem, options, or scenario.
   - Write the MCQ as a self-contained board-style question that can be answered using only the information implicitly contained in the vignette (even though it is derived from the provided rules/source text).

Validation Step:
Before finalizing the correct answer:
   - Double-check numbers against SOURCE TEXT.
   - If rule says "24 hours" but text says "2 days", use "2 days".

Grounding constraint for Explanation (MUST):
   - In the explanation, do NOT assert any medical facts, contraindications, preferences, or rationales that are NOT explicitly supported by the provided rules_json or source_fragment/source text.
   - If a distractor is incorrect because it is outside the provided rules, say so explicitly (e.g., "This is not supported by the provided source text/rules") rather than introducing outside clinical knowledge.

Output JSON format (Return JSON ONLY):

{*EXPECTED JSON FORMAT*}

Prompt A4. Profile-Question Core Decomposition

[**Prompt**]

You are a clinical information extraction assistant.
Your task is to separate a clinical question into two parts:
1). PATIENT_PROFILE - Basic demographic information about the patient, such as age, sex, or weight. Do not include any clinical information in this part, such as clinical history or lab values.
2). QUESTION_CORE - The remaining clinical question content describing the medical situation, condition, procedure, or decision.
Return the output strictly in following JSON format without any explanation and reasoning.
{
"PATIENT_PROFILE": "extracted patient profile",
"QUESTION_CORE": "extracted question core"
}
If no patient profile exists, return:
PATIENT_PROFILE = "general patient"
QUESTION_CORE = original question

Clinical Question: {*q*}

Output:

Prompt A5. Base QA Prompt for Our Benchmark

[**Prompt**]

Instruction: The following is a scenario-based multiple-choice question for anesthesiology specialists. Carefully consider the patient profile and clinical scenario, and choose the most accurate answer from the options provided. Be sure to follow the output structure and provide an appropriate explanation.
Output structure:

- Explanation: Your step-by-step reasoning
- Final answer: Your final answer (e.g. (A), (B), ...)

Patient profile: {*patient_profile*}
Clinical scenario: {*scenario*}
Question: {*question*}
{*option_str*}

Answer:

Prompt A6. Base QA Prompt for other benchmarks

[**Prompt**]

Instruction: The following is a scenario-based multiple-choice question for anesthesiology specialists. Carefully consider the patient profile and clinical scenario, and choose the most accurate answer from the options provided. Be sure to follow the output structure and provide an appropriate explanation.
Output structure:

- Explanation: Your step-by-step reasoning
- Final answer: Your final answer (e.g. (A), (B), ...)

Question: {*question*}
{*option_str*}

Answer:

Prompt A7. QA Prompt with RAG for Our Benchmark

[**Prompt**]

```
### Instruction: The following is a scenario-based multiple-choice question for anesthesiology
specialists. Carefully consider the patient profile and clinical scenario, and choose the most
accurate answer from the options provided. If the retrieved excerpts are included, use them as
supporting evidence to help identify the correct answer when relevant. Be sure to follow the
output structure and provide an appropriate explanation.
### Output structure:
    • Explanation: Your step-by-step reasoning
    • Final answer: Your final answer (e.g. (A), (B), ...)

### Patient profile: {patient_profile}
### Clinical scenario: {scenario}
### Question: {question}
{option_str}

### Retrieved excerpt (top1)
[1] Text: {retrieved_context}
### Answer:
```

Prompt A8. QA Prompt with RAG for other benchmarks

[**Prompt**]

```
### Instruction: The following is a scenario-based multiple-choice question for anesthesiology
specialists. Carefully consider the patient profile and clinical scenario, and choose the most
accurate answer from the options provided. If the retrieved excerpts are included, use them as
supporting evidence to help identify the correct answer when relevant. Be sure to follow the
output structure and provide an appropriate explanation.
### Output structure:
    • Explanation: Your step-by-step reasoning
    • Final answer: Your final answer (e.g. (A), (B), ...)

### Question: {question}
{option_str}

### Retrieved excerpt (top-1)
[1] Text: {retrieved_context}
### Answer:
```

Retrieval Case Analysis

Case 1: Pediatric intraoperative fluid selection

Question: A 6-year-old girl weighing 20 kg is scheduled for an elective orthopedic procedure under general anesthesia. She has been fasting according to standard NPO guidelines. During the intraoperative period, the anesthesia provider needs to select an intravenous fluid for maintenance requirements and to compensate for volume deficits related to her fasting state. Which of the following intravenous fluid strategies is most appropriate for this patient to minimize the risk of acute electrolyte complications?
(A): Administration of 0.45% NaCl (half-normal saline)
(B): Administration of 5% dextrose in water (D5W)
(C): Administration of 0.18% NaCl with 4% glucose
(D): Administration of an isotonic crystalloid solution
Answer: (D)

Base-RAG retrieved document: Intravenous fluid administration has been occurring for well over one hundred years for a variety of pediatric disease processes. Surprisingly, clinicians have yet to agree upon a standardized intravenous solution. There is ongoing debate regarding the administration of isotonic versus hypotonic fluids as maintenance solutions. In this article, we will review what is known about different maintenance solutions, discuss the potential complications with their use, and summarize the available evidence to help guide clinicians in their choice of maintenance fluids for their patients.

Base-RAG answer: (C) 0.18% NaCl with 4% glucose:** This solution is specifically designed for pediatric maintenance fluid therapy. It provides some sodium to help meet daily requirements and a small amount of glucose for energy, while being relatively isotonic. It's formulated to more closely match the electrolyte content of plasma and minimize the risk of hyponatremia or hypernatremia.**(D) Isotonic crystalloid solution:** While isotonic crystalloids (like normal saline or Lactated Ringer's) are good for initial volume resuscitation, they are generally *more* concentrated in sodium than plasma and can lead to hypernatremia if used as a sole maintenance fluid, especially with limited renal function. **The retrieved excerpt highlights the ongoing debate regarding isotonic vs. hypotonic fluids, but doesn't directly favor one over the other. However, considering the patient's age and the need to avoid electrolyte imbalances, a solution specifically formulated for pediatric maintenance is the safest choice. Final answer: (C)**

Prof-RAG retrieved document: This consensus- based S1 Guideline for perioperative infusion therapy in children is focused on safety and efficacy. The objective is to maintain or re-establish the child's normal physiological state (normovolemia, normal tissue perfusion, normal metabolic function, normal acid- base-electrolyte status). Therefore, the perioperative fasting times should be as short as possible to prevent patient discomfort, dehydration, and ketoacidosis. A physiologically composed balanced isotonic electrolyte solution (BS) with 1-2.5% glucose is recommended for the intraoperative background infusion to maintain normal glucose concentrations and to avoid hyponatremia, hyperchloremia, and lipolysis. Additional BS without glucose can be used in patients with circulatory instability until the desired effect is achieved. The additional use of colloids (albumin, gelatin, hydroxyethyl starch) is recommended to recover normovolemia and to avoid fluid overload when crystalloids alone are not sufficient and blood products are not indicated. Monitoring should be extended in cases with major surgery, and autotransfusion maneuvers should be performed to assess fluid responsiveness.

Prof-RAG answer: The question asks for ... Let's analyze each option: ... **(C) 0.18% NaCl with 4% glucose:** This option is a good contender as it provides some sodium and glucose. However, the sodium concentration is quite low, and it still doesn't offer a full electrolyte profile. **(D) An isotonic crystalloid solution:** Isotonic crystalloids (like Lactated Ringer's or Plasma-Lyte) provide both fluid and electrolytes in a concentration similar to blood, making them ideal for maintaining volume and electrolyte balance. **The retrieved excerpt [1] specifically recommends a "physiologically composed balanced isotonic electrolyte solution (BS) with 1-2.5% glucose" for intraoperative infusion. This supports the use of an isotonic, balanced crystalloid solution. The excerpt also highlights the importance of avoiding hyponatremia and hyperchloremia, which isotonic balanced solutions help prevent. Therefore, an isotonic crystalloid solution is the most appropriate choice to minimize the risk of acute electrolyte complications while addressing the patient's maintenance fluid and fasting-related volume deficits. Final answer: (D)**

Figure A3: Case 1 – Full example of the success case summarized in Figure 4. Prof-RAG retrieves a pediatric perioperative guideline, enabling correct fluid selection.

Case 2: Pediatric postoperative fluid rate selection

Question: A 6-year-old child weighing 20 kg is in the postoperative period following an abdominal procedure. Parenteral fluid management is necessary. The patient's hourly urine volume is diminished, attributed to postoperative ADH secretion. Which of the following is the most appropriate parenteral fluid management strategy?
(A): Isotonic solution at 30 ml/h (B): Isotonic solution with 5% glucose at 60 ml/h
(C): Isotonic solution at 60 ml/h (D): Hypotonic solution at 30 ml/h
Answer: (A)

Base-RAG retrieved document: The entire team (including anesthesiologists, surgeons, and intensive care physicians) must work together to determine the optimal amount and type of fluid necessary in the perioperative period. The authors present an overview of the basic principles that underlie fluid management, including evidence-based recommendations and a rational approach for when and what to administer.

Base-RAG answer: **(C) Isotonic solution at 60 ml/h:** This provides a reasonable maintenance fluid rate for a 20 kg child (approximately 3 ml/kg/hr, within the typical maintenance range). Isotonic solutions do not contribute to electrolyte imbalances and are generally well-tolerated. **(D) Hypotonic solution at 30 ml/h:** A hypotonic solution would worsen any potential hyponatremia caused by ADH secretion. **The most appropriate approach is to provide maintenance fluids with an isotonic solution at a rate that addresses the patient's needs without contributing to fluid overload. Final answer: (C)**

Prof-RAG retrieved document: Maintenance fluids following major operations in children are typically administered with a continuous rate. We hypothesized that administering fluids as intermittent boluses is more physiologic and could limit post-operative fluid volume, thereby avoiding harmful effects of excess fluid. [...] Bolus group patients received significantly less fluid up to 72 h post-operatively with average rates of 0.49 mL/kg/h vs 0.86 mL/kg/h at 0–24 h ($p \ll 0.01$). Additionally, the bolus group maintained adequate urine output, tolerated a regular diet sooner (2.08 vs 2.51 days; $p = 0.0023$) and averaged a shorter hospital stay (3.12 vs 4.14 days; $p = 0.004$).

Prof-RAG answer: **(A) Isotonic solution at 30 ml/h:** This equates to 1.5 ml/kg/hr, a conservative rate appropriate given ADH-induced water retention. **(C) Isotonic solution at 60 ml/h:** This equates to 3 ml/kg/hr, which could worsen fluid overload in the presence of ADH. **The excerpt highlights that reducing fluid volume post-surgery leads to faster recovery. While it does not directly address ADH secretion, it supports conservative fluid administration. Considering the patient's condition, a lower rate of isotonic solution is most appropriate. Final answer: (A)**

Figure A4: Case 2 – Pediatric postoperative fluid rate management. Additional success example. Base-RAG retrieved a general overview of perioperative fluid principles lacking specific rate guidance, leading the model to select the standard maintenance rate (60 ml/h) without accounting for ADH-mediated fluid retention. Prof-RAG retrieved a pediatric surgical study demonstrating benefits of conservative fluid administration, which supported the reduced rate (30 ml/h) consistent with the '2-1-0.5' rule for diminished urine output.

Case 3: Pediatric IPAH with RV failure — risk assessment and management
Question: A 7-year-old child (22 kg) with idiopathic pulmonary arterial hypertension (IPAH) presenting for preoperative evaluation prior to an elective umbilical hernia repair. The patient has a history of recurrent syncope and is currently classified as WHO functional class IV. Physical examination reveals clinical evidence of right ventricular (RV) failure. Preoperative echocardiography demonstrates significant RV enlargement and a small pericardial effusion. Recent cardiac catheterization data shows a cardiac index (CI) of 1.8 $\mathrm{L/min/m^2}$ and a pulmonary vascular resistance to systemic vascular resistance (PVR/SVR) ratio of 1.1. Which of the following is the most appropriate assessment of this patient's risk and management?
(A): The patient is classified as lower risk; diltiazem is indicated if the heart rate is elevated.
(B): The patient is classified as higher risk; nifedipine is indicated if acute vasodilator testing is positive.
(C): The patient is classified as higher risk; calcium channel blockers are contraindicated.
(D): The patient is classified as lower risk; verapamil is the preferred agent for its pulmonary vasoreactive properties.
Answer: (C)

Base-RAG retrieved document: Pulmonary arterial hypertension (PAH) is an uncommon disease characterized by progressive remodeling of the distal pulmonary arteries, resulting in elevated pulmonary vascular resistance and, eventually, in right ventricular failure. A thorough evaluation of a patient is critical to correctly characterize the PH. Given the multitude of treatment options currently available for PAH, assessment of risk and response to therapy is critical in long-term management. This review also underscores unique situations, including perioperative management, intensive care unit management, and pregnancy, and highlights the importance of collaborative care of the PAH patient through a multidisciplinary approach.

Base-RAG answer: **(C) Higher risk; calcium channel blockers are contraindicated:** The patient is classified as WHO functional class IV, with RV failure, significant RV enlargement, a low cardiac index (1.8 $\mathrm{L/min/m^2}$), and a PVR/SVR ratio of 1.1, all supporting higher risk classification. CCBs can be harmful in patients who do not respond to vasodilator testing, as they can lead to systemic vasodilation, potentially worsening right ventricular failure. Given the patient's severe symptoms and evidence of RV failure, option (C) is the most appropriate assessment.
Final answer: (C)

Prof-RAG retrieved document: Multivariate analysis revealed only a baseline PVR-to-systemic vascular resistance ratio of 0.8 or greater as a significant predictor of late severe PAH. Conclusions: long-term follow-up demonstrated that 60% of the patients will achieve normal or nearly normal pulmonary artery pressures. Furthermore, the study demonstrated that sildenafil improves preoperative hemodynamics and postoperative management. Children with VSD, elevated PVR, and PAH should not be denied operation.

Prof-RAG answer: **(B) Higher risk; nifedipine is indicated if acute vasodilator testing is positive:** The patient is at higher risk given WHO class IV, RV failure, and significant hemodynamic alterations. The use of calcium channel blockers depends on the response to acute vasodilator testing. Patients who respond may benefit from CCBs. **(C) Higher risk; CCBs are contraindicated:** While the patient is indeed at higher risk, the blanket statement that CCBs are contraindicated is not accurate without considering the response to acute vasodilator testing. The management should consider vasodilator testing to determine the appropriateness of CCBs like nifedipine.
Final answer: (B)

Figure A5: Case 3 – Full example of the failure case summarized in Figure 4. Prof-RAG retrieves a VSD study instead of an IPAH guideline, leading to an unsafe recommendation.

Case 4: Pediatric caudal block — alternative approach selection
Question: A 4-year-old, 18 kg male is scheduled for an elective inguinal hernia repair under general anesthesia with a planned caudal block for postoperative analgesia. Following the induction of general anesthesia, the patient is placed in the lateral decubitus position. Upon palpation of the sacral region, the anesthesia provider finds it difficult to identify the sacral hiatus. Multiple attempts to locate the landmarks for a standard hiatal approach are unsuccessful. Which of the following is the most appropriate next step for performing the regional block in this patient?
(A): Perform a trans-sacral approach at the S2-S3 level, inserting the needle at a 30-degree angle to the skin.
(B): Perform a trans-sacral approach at the S1-S2 level, inserting the needle at a 45-degree angle cephalad.
(C): Continue the hiatal approach but increase the local anesthetic volume by 50% to ensure adequate spread.
(D): Perform a trans-sacral approach at the S2-S3 level, inserting the needle perpendicularly to the skin.
Answer: (D)

Base-RAG retrieved document: Caudal epidural block is a commonly used technique for surgical anesthesia in children and chronic pain management in adults. It is performed by inserting a needle through the sacral hiatus to gain entrance into the sacral epidural space. This high failure rate could be attributed to anatomic variations that make locating sacral hiatus difficult. With the advent of fluoroscopy and ultrasound in guiding needle placement, the success rate of caudal epidural block has been markedly improved. The relevant anatomy and techniques in performing the caudal epidural block will be briefly reviewed in this article.

Base-RAG answer: **(D) Trans-sacral approach at S2-S3 with perpendicular insertion:** This technique can be used when the sacral hiatus is difficult to identify, as it allows for a more direct approach to the epidural space. The retrieved excerpt discusses anatomic variations that make hiatal identification difficult and reviews relevant anatomy and techniques. Since the options do not include ultrasound guidance, option (D) aligns with a method that could be used in challenging cases.
Final answer: (D)

Prof-RAG retrieved document: An 8-year-old boy was scheduled for urethral fistula correction under general anaesthesia with an ultrasound-guided caudal block. During scanning of the caudal area, we noticed two cystic structures in the caudal space in the region of the intended needle endpoint so we decided against performing the caudal block. Surgery was conducted uneventfully and a penile block was administered as an alternative for analgesia postoperatively. Radiological evaluation of the child 3 weeks later revealed the presence of perineural cysts in the sacral region.

Prof-RAG answer: **(B) Trans-sacral approach at S1-S2 with 45-degree angle cephalad:** This approach aligns with the anatomical considerations for accessing the caudal space when the hiatal approach is not possible. This approach and angle are designed to facilitate the needle's entry into the caudal space while minimizing the risk of complications. **(D) Trans-sacral approach at S2-S3 with perpendicular insertion:** While this is a viable alternative, the retrieved excerpt emphasizes the need for careful anatomical consideration, and the S1-S2 level with angled insertion may be more appropriate.
Final answer: (B)

Figure A6: Case 4 – Pediatric caudal block alternative approach. Additional failure example. Base-RAG retrieved a technique-focused review of caudal epidural anatomy, providing sufficient context for the model to identify the perpendicular trans-sacral approach at S2-S3. Prof-RAG retrieved a pediatric case report about perineural cysts—demographically matched but procedurally irrelevant—causing the model to fabricate an anatomical rationale for the incorrect S1-S2 level with angled insertion. This illustrates how profile-aware retrieval can overemphasize patient demographics at the expense of procedural question intent.

Annotation Guidelines

Open Benchmark Question Categorization Guideline

[Annotation Guideline]

Purpose: The goal of this annotation is to classify each question (with multiple-choice options and answer) according to:

- Question Type
- Target Population
- Anesthesiology Relevance

Annotators should rely on the content of the question, options, and the answer.

Question Type Classification

Select one of the following categories:

MANAGEMENT: Questions that ask what action to take.

Examples include:

- treatment selection
- next step in management
- medication dosing
- diagnostic or therapeutic procedures
- guideline-based decisions

DIAGNOSIS: Questions that ask to identify a disease or cause.

Examples include:

- most likely diagnosis
- underlying cause
- differential diagnosis

BASIC_SCIENCE: Questions that test factual or conceptual knowledge, not clinical decision-making.

Examples include:

- physiology
- pharmacology mechanisms
- anatomy
- definitions

OTHER: Questions that do not fit the above categories.

Examples include:

- research methods
- ethics or law
- statistics
- administrative topics
- unclear or ambiguous questions

Target Population Classification

Select one of the following:

Pediatric

Choose Pediatric if:

- the question explicitly mentions a patient under 18 years old
- the condition is exclusively or predominantly pediatric
- all answer options relate to pediatric conditions

Adult

Choose Adult if:

- the question explicitly mentions a patient aged 18 or older
- the scenario refers to adults, elderly, or geriatric patients

General
Choose General if:

- no age group is specified
- the scenario applies broadly to both pediatric and adult patients
- the population is unclear or mixed

Anesthesiology Relevance
Determine whether the question is primarily about anesthesiology or pain medicine.
TRUE
Select TRUE if:

- the main subject is anesthesia or pain management
- the question would most naturally be answered by an anesthesiologist
- the focus is on anesthetic drugs, airway, perioperative management, or anesthesia-related decisions

FALSE
Select FALSE if:

- the main subject is another specialty (e.g., internal medicine, surgery, pediatrics, emergency medicine)
- anesthesia is only a minor or background detail
- the question focuses on diagnosis or general disease management rather than anesthesia

Clinical Question/Answer Quality Evaluation Guideline

[**Evaluation Guideline**]

Purpose: The goal of this annotation is to evaluate the quality of multiple-choice clinical questions generated from anesthesia guidelines.
Each question should be evaluated according to four criteria:

- Guideline Consistency
- Medical Validity
- One-Best-Answer Clarity
- Distractor Plausibility
- Target Population Appropriateness

Each item should be scored independently.
Guideline Consistency
This criterion evaluates whether the question, answer, and explanation correctly reflect the recommendations described in the source guideline.
Annotators should consider:

- Whether the recommended answer aligns with guideline recommendations
- Whether the explanation accurately reflects guideline content
- Whether the question correctly represents the clinical scenario described in the guideline

Rating Scale

- 1: Not consistent with the guideline (None of the information provided in the guideline appears in the generated QA or clinical scenario.)
- 2: Mostly inconsistent
- 3: Partially consistent
- 4: Mostly consistent
- 5: Fully consistent with guideline recommendations (Key elements from the guideline|including patient profile, clinical scenario, actions, and treatment recommendations|are appropriately reflected in the QA set.)

Medical Validity
This criterion evaluates whether the clinical reasoning and answer are medically correct, regardless of guideline wording.
Annotators should consider:

- Clinical plausibility
- Alignment with accepted medical practice
- Logical correctness of the explanation

Rating Scale
- 1: Medically incorrect (The core question or answer contains clinically incorrect or unsafe information.)
- 2: Mostly incorrect
- 3: Uncertain / partially valid
- 4: Mostly correct
- 5: Clearly medically correct (The question and answer are clinically correct and the patient profile and scenario are medically plausible.)

One-Best-Answer Clarity
This criterion evaluates whether the question clearly has one correct answer.
Annotators should consider:
- Whether multiple options could reasonably be correct
- Whether the correct answer is clearly the best choice
- Whether the question wording creates ambiguity

Rating Scale
- 1: Multiple answers appear correct (Multiple options appear correct or the correct answer is unclear.)
- 2: Several options are plausible answers
- 3: Some ambiguity exists
- 4: Mostly clear single best answer
- 5: Clear and unambiguous single best answer (There is a clearly identifiable single best answer with no reasonable ambiguity.)

Distractor Plausibility
This criterion evaluates whether the incorrect answer options (distractors) are plausible and realistic.
Good distractors should:
- Be clinically plausible
- Reflect common misconceptions or alternative strategies
- Not be obviously incorrect

Rating Scale
- 1: Distractors are clearly unrealistic (Incorrect options are obviously wrong or clinically implausible.)
- 2: Mostly implausible distractors
- 3: Somewhat plausible
- 4: Mostly plausible
- 5: Highly plausible distractors (All options appear clinically plausible, and at least two distractors could reasonably confuse clinicians.)

Target Population Appropriateness
This criterion evaluates whether the target population assigned to each question is appropriate.
Each question has a predefined population label:
- Pediatric
- Adult
- General

Annotators should determine whether the assigned label correctly reflects the clinical scenario.
Target Population Definitions
Pediatric
Select ‘‘Correct’’ if the scenario clearly refers to pediatric patients (e.g., neonates, infants, children, or adolescents). Indicators may include explicit age references, pediatric surgical procedures, or pediatric physiological context.
Adult
Select ‘‘Correct’’ if the scenario clearly refers to adult patients. Indicators may include explicit adult ages, adult-specific diseases or surgeries, or clinical contexts rarely encountered in pediatrics.
General
Select ‘‘Correct’’ if no specific age group is mentioned and the recommendation could reasonably apply to both pediatric and adult patients.
Annotation Scheme

- Correct: The assigned population label matches the clinical scenario.
- Incorrect: The assigned label does not match the clinical scenario.