Expert-Guided Prompting and Retrieval-Augmented Generation for Emergency Medical Service Question Answering

Xueren Ge[1], **Sahil Murtaza**[1], **Anthony Cortez**[2], **Homa Alemzadeh**[1]

[1]School of Engineering and Applied Sciences, University of Virginia
[2]School of Medicine, University of Virginia
zar8jw@virginia.edu, vpn9ej@virginia.edu, aec3gp@virginia.edu, ha4d@virginia.edu

Abstract

Large language models (LLMs) have shown promise in medical question answering, yet they often overlook the domain-specific expertise that professionals depend on-such as the clinical subject areas (e.g., trauma, airway) and the certification level (e.g., EMT, Paramedic). Existing approaches typically apply general-purpose prompting or retrieval strategies without leveraging this structured context, limiting performance in high-stakes settings. We address this gap with EMSQA, an 24.3K-question multiple-choice dataset spanning 10 clinical subject areas and 4 certification levels, accompanied by curated, subject area-aligned knowledge bases (40K documents and 2M tokens). Building on EMSQA, we introduce (i) Expert-CoT, a prompting strategy that conditions chain-of-thought (CoT) reasoning on specific clinical subject area and certification level, and (ii) ExpertRAG, a retrieval-augmented generation pipeline that grounds responses in subject area-aligned documents and real-world patient data. Experiments on 4 LLMs show that Expert-CoT improves up to 2.05% over vanilla CoT prompting. Additionally, combining Expert-CoT with ExpertRAG yields up to a 4.59% accuracy gain over standard RAG baselines. Notably, the 32B expertise-augmented LLMs pass all the computer-adaptive EMS certification simulation exams.

Code & Data — https://uva-dsa.github.io/EMSQA

Introduction

The rapid advancement of large language models (LLMs) has brought new possibilities to high-stakes domains such as emergency medical services (EMS) (Weerasinghe et al. 2024), where accurate and reliable decision-making is critical. There is growing interest in leveraging LLMs for medical education (Abd-Alrazaq et al. 2023), decision support (Ge et al. 2024; Luo et al. 2025), and certification preparation (Kung et al. 2023), particularly in the context of open-domain multiple-choice question answering (MCQA). However, while LLMs have shown promising performance on general medical QA benchmarks (Cai et al. 2024), important gaps remain between their current capabilities and the reasoning processes used by trained medical professionals.

Recent approaches in medical MCQA, such as chain-of-thought prompting (CoT) (Wei et al. 2022) and retrieval-augmented generation (RAG) (Lewis et al. 2020), have improved LLM performance by enhancing reasoning capabilities and incorporating external domain knowledge during inference. But these approaches often treat both reasoning and retrieval as undifferentiated processes: the model sees a question and retrieves documents or reasons directly, without considering what kind of knowledge is relevant or how a human expert would approach the task. In contrast, real-world medical professionals typically begin by identifying the subject area of the question—e.g., whether it pertains to trauma, airway management, or pharmacology—and then reason from that domain-specific perspective, using knowledge and protocols appropriate to their level of certification.

Existing benchmarks such as MedQA (Jin et al. 2021), MMLU-Med (Hendrycks et al. 2020) and MedMCQA (Pal, Umapathi, and Sankarasubbu 2022) lack both this structured representation of question expertise (e.g., subject area, certification level) and the associated domain-specific knowledge. This makes it difficult to align retrieval and reasoning processes with human-like problem-solving strategies. In particular, publicly available EMS question answering datasets and knowledge sources with expertise annotation are scarce. Further, current state-of-the-art (SOTA) methods do not account for how incorporating structured expertise can improve the overall effectiveness of reasoning and answer generation in retrieval-augmented systems.

To address these gaps, we propose a new dataset and a *domain-expertise-guided* LLM framework that models medical MCQA in a more structured, cognitively informed manner. As shown in Figure 1, our contributions are three-fold:

- We introduce **EMSQA**, the first EMS MCQA dataset of 24.3K questions, curated based on public and private sources, covering 10 subject areas and 4 certification levels, and accompanied by a structured, subject area-aligned EMS knowledge base (KB) with 40K documents and 4M real-world patient care reports. Partial data (from public sources) and the whole EMS KB is shared as a resource with the EMS and research communities.
- We propose a *expertise-guided* LLM framework that infers the domain expertise attributes and injects them into LLMs using two approaches: (1) an **expertise-guided prompting strategy (Expert-CoT)** that encour-

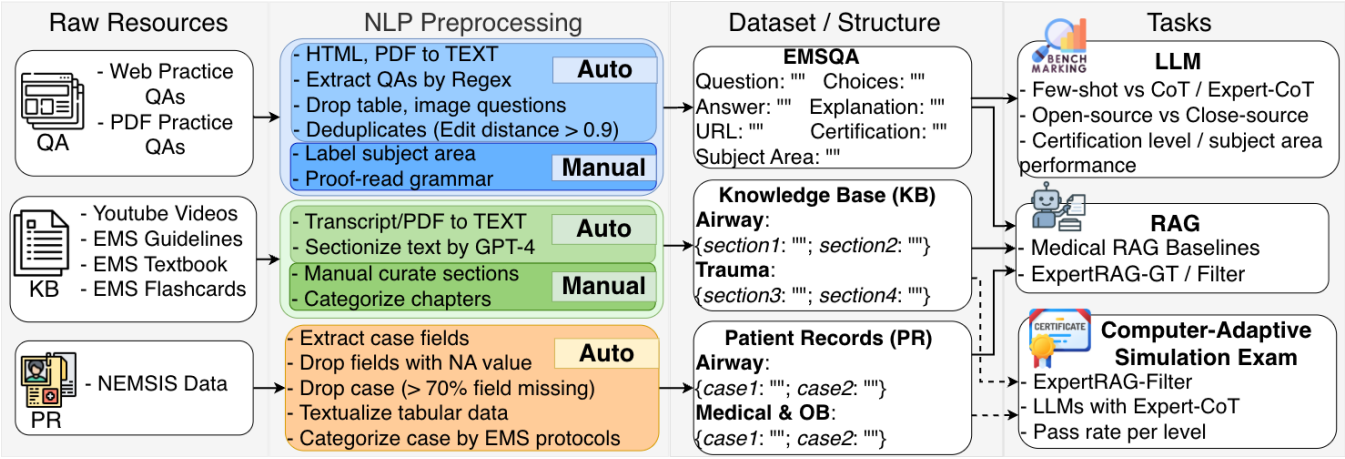

Figure 1: Overall Approach. (1) EMSQA, KB, PR Construction; (2) Tasks: Benchmark LLMs, RAG and Certification Exams.

ages step-by-step reasoning from a domain-specific perspective. (2) an **expertise-guided RAG method (ExpertRAG)** that selectively retrieves expertise-aligned knowledge from curated EMS KBs and patient records.

- We benchmark multiple LLMs on EMSQA, evaluating performance across certification levels and subject areas, and compare our framework against SOTA RAG methods. Experimental results show that combining **Expert-CoT** and **ExpertRAG** yields up to a 4.59% improvement in accuracy. Notably, the 32B expertise-augmented models pass all the EMS certification simulation exams.

Related Works

Medical Question Answering Datasets

Medical QA datasets typically fall into two paradigms: retrieval-based tasks (Pampari et al. 2018; Krithara et al. 2023; Ben Abacha and Demner-Fushman 2019), which require explicit evidence grounding by locating answers within documents, and open-domain multiple-choice tasks such as MedQA (Jin et al. 2021), MMLU-Med (Hendrycks et al. 2021), and MedMCQA (Pal, Umapathi, and Sankarasubbu 2022) that test implicit medical reasoning by choosing the best answer based on world knowledge. However, existing MCQA datasets focus on a single certification level and provide subject area labels without corresponding knowledge bases. EMSQA is the first open-domain medical MCQA dataset that spans multiple certification levels while also furnishing both subject area annotations and a structured EMS knowledge base. Table 1 shows a comparison between EMSQA and SOTA open-domain MCQA datasets.

Retrieval Augmentation Generation

The basic RAG framework (Lewis et al. 2020) couples a seq2seq model with a dense Wikipedia retriever, and has since been improved via query rewriting (Chan et al. 2024), entity graphs (Edge et al. 2024), and document-structure–aware retrieval (Li et al. 2025a), though these methods mainly target general-domain text. For the medical domain, MedRAG (Xiong et al. 2024a) proposes a RAG pipeline with hybrid sparse–dense retrieval on MedCorps, and i-MedRAG (Xiong et al. 2024b) extends it with follow-up question generation and interactive reasoning. Self-BioRAG (Jeong et al. 2024) adapts Self-RAG (Asai et al. 2023) with domain-specific retrieval triggers, while RAG^2 (Sohn et al. 2024) leverages rationale-based queries and filtering. EXPRAG (Ou et al. 2025) retrieves similar patient cases, and ClinicalRAG (Lu, Zhao, and Wang 2024) exploits medical entities to query corpora. However, existing medical RAG systems largely ignore question-specific expertise (e.g., subject area or certification) as explicit signals to guide retrieval and reasoning. Unlike Metadata-RAG (Oudenhove 2024), which parses metadata directly from the query, we inject expertise attributes inferred from the question by a model trained on our Q&A dataset.

	MedQA	**MedMCQA**	**EMSQA**
Domain	General Med.	General Med.	EMS
Data Size	12.7K	193K	24.3K
Exam	USMLE	AIIMS&NEET PG	NREMT
#Certification	1	1	4
#Subject Area	✗	21	10
KB	Raw	✗	Categorized

Table 1: Comparison of English medical MCQA datasets

EMSQA

Data Collection and Preprocessing

We collected a total of 24.3K multiple-choice questions and their corresponding answer choices from practice tests available on 17 websites targeting the National Registry of Emergency Medical Technicians (NREMT) examination (NREMT 2001–2025). The NREMT exam is administered as a Computer Adaptive Test (CAT), which dynamically adjusts question difficulty based on the examinee's performance, providing a personalized and efficient assessment. It certifies providers at 4 ascending levels of difficulty: entry-level, Emergency Medical Responder (EMR); basic-level,

Set	Size	Type	Criteria	vs KB	vs PR
Public	Train(13,021) Val(1,860) Test(3,721)	Semantic	Avg Sim	79.21	66.45
		Syntactic (hit rate)	Vocab	82.95	21.14
			Cpt w/o norm	41.65	8.87
			Cpt w/ norm	63.30	15.28
Private	Test(5,669)	Semantic	Avg Sim	80.75	75.35
		Syntactic (hit rate)	Vocab	90.89	28.26
			Cpt w/o norm	53.18	14.36
			Cpt w/ norm	72.49	22.66

Table 2: Statistics by split for Public and Private EMSQA and Semantic and syntactic evaluation of QA overlap vs. KB/PR. Cpt: Concept; norm: medical normalization.

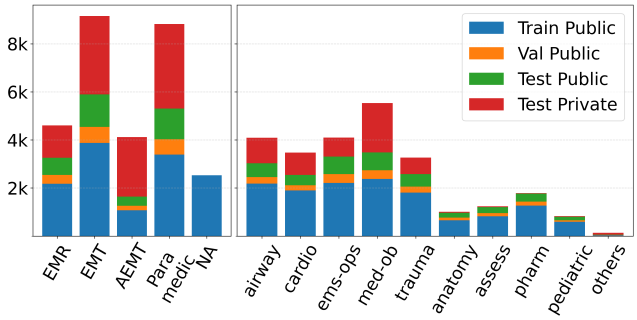

Figure 2: Distributions of Questions by Certification Level (Left) and Subject Area (Right).

Emergency Medical Technician (EMT); intermediate-level, Advanced EMT (AEMT); and advanced-level, Paramedic. Our dataset covers all four certification levels. These practice tests assess examinees' ability to apply medical knowledge, concepts, and principles, as well as their capacity to demonstrate fundamental patient-centered skills.

The overall EMSQA dataset comprises questions from both public and private (subscription-based) websites, but we only release the portion derived from public materials because of copyright restriction of private websites (See **Appendix A.2 in Extended version** for more details). To ensure the data quality, we did both automatic preprocessing and manual verification of the raw data as shown in Figure 1:

- Heuristic rules were used to extract and remove special tokens like HTML tags, special symbols.
- Each question is well-structured and represented as a dictionary containing the following fields: the question text, answer options, correct answer, explanation, source URL, certification level, and subject area.
- Questions related to images and tables were excluded. All the questions are answerable using text inputs only.
- All duplicate questions were removed by computing the Levenshtein distance between each pair of questions in the dataset. Pairs with a similarity score greater than 0.9 were considered duplicates and subsequently removed.
- All questions are manually labeled with subject areas, and both questions and answer choices have undergone human proofreading to correct any grammatical errors. A sample of 100 questions and KB documents was further verified by an EMT expert (see Appendix A.5).

Data Statistics

As shown in Table 2, the dataset includes a total of 18,602 and 5,669 practice questions from public and private sources, respectively. We use the private questions exclusively for testing and split the public questions into train, validation, and test sets of 13,021, 1,860, 3,721 questions, with average token lengths of 18.27, 19.12, 18.99, respectively. More detailed statistics are provided in Appendix A.2.2.

Figure 2 details the distribution of questions by certification level and subject area. Because certification level is used for evaluation, all questions whose certification level marked with "NA" are relegated to the training split, leaving the validation and test splits to include only questions with explicit EMS certification levels. Subject areas including "airway", "cardiology", "EMS operations", "medical&OB", and "trauma" dominate the corpus, which mirrors the five domains mandated by the NREMT examination guidelines. The remaining subject areas—"anatomy", "assessment", "pharmacology", "pediatrics", and "others"—are present as well, but they appear less frequently, reflecting their secondary coverage in the practice materials we collected.

Knowledge Collection and Preprocessing

To build an external KB for EMSQA, we curated 16 open-access EMS education resources from reputable websites. As shown in Figure 1, these resources span four media types: YouTube video transcripts (Carrie Davis 2025; The EMS Professor 2025), official EMS guidelines (ODEMSA 2025), EMS education textbooks (American Red Cross 2025) or lecture slides (Jones & Bartlett Learning 2025), and EMS flashcards (EMT-Prep 2025). We segmented each YouTube transcript into sections using the title cues supplied by the uploader. The PDF documents were converted to plain text with PyPDF2 (Fenniak et al. 2022) and split to chapters using page ranges from their tables of contents. We then leveraged GPT-4o (Achiam et al. 2023) to reorganize the raw chapter text into a coherent section-level hierarchy (See Appendix A.3.1). Finally, we manually audited each section to ensure its heading and text span aligned with the corresponding passage in the original PDF and corrected any discrepancies. Due to the lack of certification information in the sources, we only categorized the chapters into 10 subject areas based on their titles: "airway&ventilation", "anatomy", "assessment", "cardiovascular", "ems operations", "medical&ob", "pediatrics", "pharmacology", "trauma" and "others". Our final EMS KB comprises 39,652 sections, totaling 2,545,192 tokens and 34,110 unique vocabularies.

To evaluate the helpfulness of the KB for EMSQA, we assess our KB's coverage from both syntactic and semantic aspects. For semantic evaluation, we leverage MedCPT (Jin et al. 2023) to retrieve, for each question, its most similar document from the KB and report the average similarity score between each question and its retrieved document. For syntactic evaluation, we removed stop words and com-

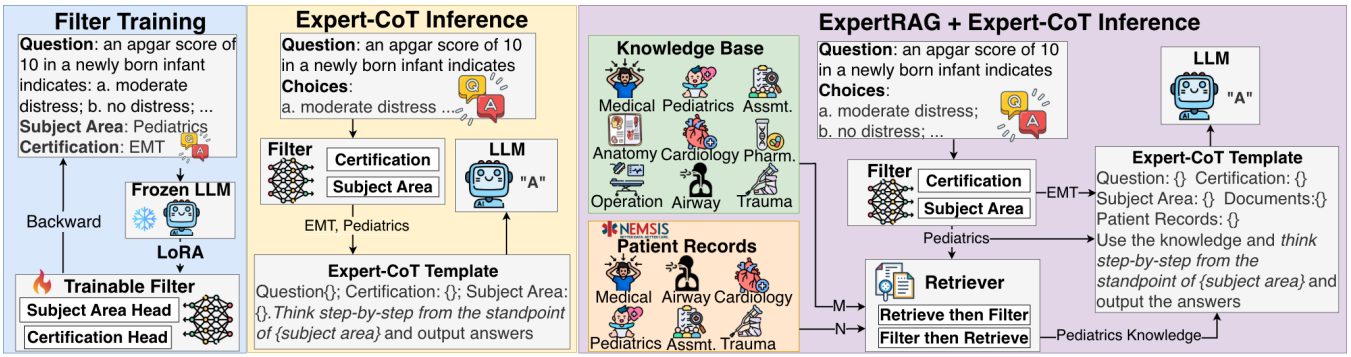

Figure 3: Expertise-Guided LLM Framework: Filter training, Expert-CoT Inference, and ExpertRAG Inference.

puted vocabulary hit rate (KB ∩ QA / QA) between the KB and EMSQA. We also follow the methods in (Ge et al. 2024) to extract EMS-related concepts defined in (Preum et al. 2020), both before and after UMLS normalization (Bodenreider 2004). Table 2 reports both syntactic and semantic overlaps with our KB. Syntactic hit rates span from 41.65% to 90.89%. Semantic similarity, measured by average cosine similarity, was 79.21% on public and 80.75% on private data. This indicates that our KB can support answering most EMSQA questions. Appendix A.3 provides comprehensive details on web crawling, EMS concept extraction, KB statistics, and the overlap between EMSQA and the KB.

Patient Care Report Collection and Preprocesing

We used the National EMS Information System (NEMSIS) (Dawson 2006) 2021 public research dataset as our source of patient records (PR). NEMSIS is a large tabular corpus in which each record includes fields such as dispatch details, scene information, initial assessment, EMS protocols and triage, vital signs, medications and procedural interventions, and patient history. As shown in Figure 1, we removed any field whose value was "NA," "Not Applicable," "Not Recorded," or "Unknown." If more than 30% of a record's fields were discarded, we excluded the entire record. Finally, we converted each remaining record into plain text by concatenating its key–value pairs and categorized the patient records into 6 subject areas based on their protocol field: "airway", "assessment", "cardiovascular", "medical&ob", "pediatrics", and "trauma". Our final corpus comprises 4,003,430 records with an average token length of 311.7. As shown in Table 2, semantic coverage of NEMSIS over EMSQA is high, with patient records reaching 66.45%/75.35% similarity on the public/private split. In contrast, syntactic concept hit rates are much lower (8.87%–28.26%) than those of the KB. Extracted NEMSIS fields and summary statistics are provided in Appendix A.4.

Methodology

Task Formulation

Given the ith question q_i and its answer options $\mathcal{O}_i = \{o_1, \ldots, o_m\}$, the goal of MCQA is to maximize the likelihood of selecting the correct answer $a_i^* \in \mathcal{O}_i$. Let $\mathcal{R}$ be a retriever that takes q_i as input and returns a set of relevant documents. A language model f selects an answer a_i as:

$$a_i = \arg\max_{o \in \mathcal{O}_i} f(o \mid q_i, \mathcal{O}_i, \mathcal{R}(q_i)) \tag{1}$$

We propose an *expertise-guided* LLM framework (see Figure 3) with an expertise classification module (called **Filter**), which infers the domain expertise attributes of subject area s_i and certification level l_i from the input question q_i, and incorporates them into f using two strategies: (1) **Expert-CoT**, a prompting strategy that encodes s_i and l_i into a prompt template to guide f based on question-specific expertise; and (2) **ExpertRAG**, a subject-area-specific $\mathcal{R}$ that retrieves knowledge sources conditioned on s_i.

Filter Design

To guide the LLM reasoning and RAG retrieval based on question-specific expertise, we train a lightweight LLM-based filter to infer the key expertise attributes, including question's subject area and certification level. As shown in Figure 3 (Left), we adopt LoRA (Hu et al. 2022) to inject a small set of trainable parameters into the model while keeping the full LLM weights fixed. We augment the LoRA modules with two classification heads, W_{sub} and W_{lvl}, that predict the question's subject area and certification level, and optimize them jointly in a multi-task setting (Li et al. 2025b). We append a special token `<classify>` at the end of each query, and extract the hidden state h_i of this final token from the last layer and feed it into our two classification heads:

$$\begin{aligned} h_i &= \text{LM}_{\text{last}}(q_i, \mathcal{O}_i \| \langle \text{classify} \rangle), \\ (p_i^{\text{sub}}, p_i^{\text{lvl}}) &= \left(\sigma(W_{\text{sub}}^\top h_i), \sigma(W_{\text{lvl}}^\top h_i)\right) \end{aligned} \tag{2}$$

where σ is sigmoid function. We use binary cross-entropy for subject area classification and cross-entropy for certification classification. We set hyper-parameters w_{sub} and w_{lvl} to balance the multi-task loss. The overall training loss is:

$$\mathcal{L} = w_{\text{sub}} \cdot \text{BCE}\left(p_i^{\text{sub}}, y_i^{\text{sub}}\right) + w_{\text{lvl}} \cdot \text{CE}\left(p_i^{\text{lvl}}, y_i^{\text{lvl}}\right) \tag{3}$$

During inference, given a question q_i with answer options $\mathcal{O}_i$, the filter first predicts a certification-level probability distribution p_i^{lvl} and a multi-label subject area probability vector p_i^{sub}. The predicted certification level and subject area set are computed as:

$$\hat{s}_i = \mathbf{1}\{p_i^{\text{sub}} > 0.5\}, \quad \hat{l}_i = \arg\max p_i^{\text{lvl}}. \tag{4}$$

Expertise-Guided Prompting (Expert-CoT)

Figure 3 (Middle) illustrates our Expert-CoT prompting method. Standard CoT prompts encourage LLMs to reason step by step but do not specify where to begin. In contrast, Expert-CoT prompting guides the model's reasoning by explicitly providing the subject area and certification level as starting point for the thought process. The final answer is generated by passing the predicted subject area $\hat{s}_i$ and certification level $\hat{l}_i$, into the Expert-CoT prompt template:

$$\hat{A}_i = f^{\text{CoT-Expert}}\big(q_i,\ \mathcal{O}_i,\ \hat{l}_i,\ \hat{s}_i\big). \tag{5}$$

Expertise-Guided RAG (ExpertRAG)

As shown in Figure 3 (Right), for ExpertRAG the filter's predicted subject area guides the retriever to search for relevant knowledge base entries and patient records tailored to the question's subject area. The LLM then conditions on the predicted expertise and the retrieved documents to generate the final answer. Based on q_i and the predicted subject area $\hat{s}_i$, we explore three retrieval strategies for retriever $\mathcal{R}$:

- **Global**: Retrieve the top M and N evidence documents from the entire KB and PR, respectively. This serves as a baseline corresponding to standard RAG without any subject area filtering.
- **Filter then Retrieve (FTR)**: First filter the whole KB and PR to retain only documents matching the predicted subject area $\hat{s}_i$, then retrieve the top M and N documents from these filtered subsets.
- **Retrieve then Filter (RTF)**: First retrieve a larger candidate set from the whole KB and PR (e.g., $10 \times M$ from KB and $10 \times N$ from PR), then filter out documents whose subject area do not match $\hat{s}_i$, retaining the top M and N relevant documents.

The final answer is generated by passing the retrieved documents, along with the predicted subject area and certification level, into the RAG prompt template:

$$\hat{A}_i = f^{\text{RAG}}\big(q_i,\ \mathcal{O}_i,\ \mathcal{R}(q_i, \hat{s}_i),\ \hat{l}_i,\ \hat{s}_i\big). \tag{6}$$

Experiments

We conduct extensive experiments to evaluate Expert-CoT and ExpertRAG methods by applying them to different baseline LLMs and comparing their performance to SOTA LLMs and RAGs. We aim to answer three research questions:
RQ1: Where do SOTA LLMs shine or stumble on EMSQA across subject areas and certification levels?
RQ2: How much does explicit expertise injected by Expert-CoT and ExpertRAG lift baseline accuracy?
RQ3: Can expertise-aware LLMs pass the NREMT standardized tests at different certification levels?

LLM Baselines

We use three categories of SOTA baseline models: (1) **Open-source LLMs**: we select Qwen3-32B (Team 2025), and LLama-3.3-70B (Grattafiori et al. 2024) because of their great performance in multiple domains; (2) **Medical LLMs**: we select OpenBioLLM-70B (Ankit Pal 2024), which currently leads the Open Medical-LLM Leaderboard (Pal et al. 2024). (3) **Closed-source LLMs**: we choose OpenAI-o3 (Brown et al. 2020) and Gemini-2.5-pro (Team et al. 2023), both of which achieve top results on a range of benchmarks. We further apply 0- to 64-shot, CoT (Wei et al. 2022), and Expert-CoT prompting to each baseline to benchmark the performance across prompt strategies.

RAG Baselines

We select the following SOTA medical RAG models due to their superior performance and code availability: (1) MedRAG (Xiong et al. 2024a), which is a RAG toolkit combining multiple medical documents; (2) i-MedRAG (Xiong et al. 2024b), which iteratively refines medical queries via multi-step retrieval; (3) Self-BioRAG (Jeong et al. 2024), which self-reflectively decides when to retrieve biomedical texts and then generates the answer; (4) Qwen3-4B + KB, which is a vanilla RAG pipeline with our collected KB as retrieval corpora; (5) Qwen3-4B + PR, a vanilla RAG pipeline with our collected PR as retrieval corpora; (6) Qwen3-4B + Global, a vanilla RAG pipeline with both KB and PR as retrieval corpora. We also include (7) Qwen3-4B with 0-shot and (8) Qwen3-4B with CoT prompting as baselines. For a fair comparison, we applied RAG with Qwen3-4B across all methods, except Self-BioRAG, which is trained from scratch. All baseline RAG methods used CoT prompting.

Implementation Details

For Expert-RAG, we use Qwen3 as the core LLM, as it is the best-performing open-source model in our benchmarking. We employ MedCPT as our retriever due to its strong performance in medical domain and widespread use in SOTA RAG. We fix the number of retrieved documents at $M = 32$ for KB retrieval and $N = 8$ for PR retrieval. All KB and PR documents are chunked with a window of 512 tokens with an overlap of 128 tokens. Since some questions in EMSQA have multiple correct answers, we report both exact-match accuracy (Acc) and sample-based F1 (Khashabi et al. 2018).

ßTo train the filter, we fine-tune LoRA modules with rank r=8, scaling factor α=16, and a dropout rate of 0.05, using the sequence length of 128 tokens. To balance the certification and subject area classification objectives, we apply DWA (Liu, Johns, and Davison 2019) with $T = 2$ to dynamically adjust w_{cat} and w_{lvl}. We use AdamW (Loshchilov and Hutter 2017) optimizer and regularization with a weight decay of 0.01. We set the decision threshold of 0.5 for subject area classification. We fix the random seed at 42, and run all experiments on NVIDIA H200 GPUs.

Experimental Results

LLM Benchmarking

To evaluate the overall performance of SOTA LLMs on EMSQA (**RQ1**), we benchmark multiple LLMs under different prompting strategies. Figure 4 and Table 3 present the results. We highlight several key findings from the evaluation:

Closed-source models outperform open-source models. In particular, OpenAI-o3 consistently achieves the highest overall accuracy of 92.39. Among open-source models,

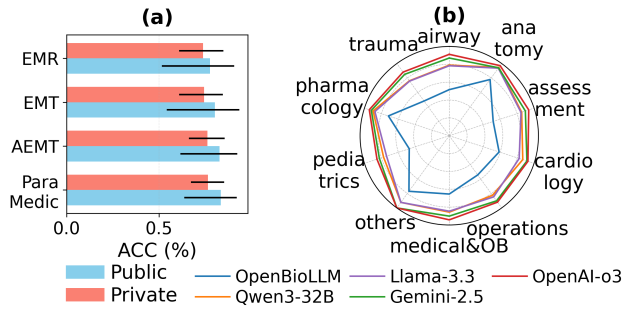

Figure 4: **(a)** Certification level 0-shot performance. **(b)** Subject area 0-shot performance on the Public dataset

Model	Prompt	Public		Private	
		Acc	F1	Acc	F1
OpenBioLLM (GT)	0-shot	57.67	57.76	63.86	64.76
	CoT	59.88	60.34	67.01	67.77
	Expert-CoT	**61.92**	**62.03**	**68.75**	**69.82**
(Filter)	Expert-CoT	61.32	61.93	67.79	68.32
Llama-3.3 (GT)	0-shot	81.69	82.69	78.06	78.77
	CoT	81.89	83.08	85.16	86.35
	Expert-CoT	**82.42**	**83.35**	86.49	87.62
(Filter)	Expert-CoT	82.40	83.18	**86.63**	**87.65**
Qwen3-32B (GT)	0-shot	83.55	83.55	85.11	85.89
	4-shot	84.41	84.41	85.48	86.13
	32-shot	81.13	81.13	82.22	83.41
	64-shot	82.48	82.48	86.22	87.26
	CoT	84.96	84.97	88.78	90.13
	Expert-CoT	**85.70**	**85.71**	**89.73**	90.98
(Filter)	Expert-CoT	85.57	85.60	89.50	**91.20**
OpenAI-o3	0-shot	**92.39**	**92.39**	–	–
Gemini-2.5	0-shot	89.36	89.36	–	–

Table 3: Accuracy and F1 (%) of LLMs under Public vs. Private Data. GT/Filter: Ground-truth/predicted expertise.

Split	Model	Method	Subject Area		Certification	
			miF	maF	miF	maF
Public	Filter	LoRA	**80.72**	**71.92**	**65.87**	**63.45**
	Qwen3-4B	0-shot	55.43	51.61	45.77	30.49
	Qwen3-4B	4-shot	56.33	54.42	45.46	30.28
	Qwen3-4B	CoT	59.72	55.66	47.80	35.77
Private	Filter	LoRA	**79.06**	**70.48**	**65.54**	**63.50**
	Qwen3-4B	0-shot	42.93	31.73	44.12	25.01
	Qwen3-4B	4-shot	45.76	34.08	44.04	29.41
	Qwen3-4B	CoT	46.22	35.49	47.70	31.92

Table 4: Expertise Classification Performance

Qwen3-32B achieves the best accuracy of 85.70, though a significant gap remains compared to closed-source models.

Few-shot prompting improves accuracy up to a point. We varied the number of in-context exemplars from 0 to 64 (See Appendix A.7.2) and observed that incorporating

Model	Description	Public		Private	
		Acc	F1	Acc	F1
No-RAG Baselines					
Qwen3-4B	0-shot	70.99	71.01	69.88	69.95
Qwen3-4B	CoT	72.35	73.09	70.58	72.02
RAG Baselines + CoT					
MedRAG	RAG on Med	74.31	74.41	71.12	73.33
i-MedRAG	Iterative RAG	77.96	78.00	74.02	76.35
Self-BioRAG	SelfRAG on Bio	55.71	58.84	45.72	49.67
Qwen3-4B	KB	76.49	76.07	75.02	76.53
Qwen3-4B	PR	73.02	73.96	70.54	72.38
Qwen3-4B	Global	78.12	79.17	75.46	76.87
RAG Baselines + Expert-CoT		$\Delta_{\text{Acc/F1}} = +1.38/+0.46$			
Qwen3-4B	KB	78.02	79.04	76.01	76.25
Qwen3-4B	PR	73.82	73.82	71.53	72.96
Qwen3-4B	Global	**79.59**	**79.61**	**76.75**	**77.35**
ExpertRAG-GT + CoT		$\Delta_{\text{Acc/F1}} = +3.35/+2.71$			
ExpertRAG	FTR	80.97	81.34	79.13	80.00
ExpertRAG	RTF	**81.11**	**81.45**	**79.17**	**80.01**
ExpertRAG-GT + Expert-CoT		$\Delta_{\text{Acc/F1}} = +4.59/+3.69$			
ExpertRAG	FTR	81.62	81.65	80.40	81.02
ExpertRAG	RTF	**82.24**	**82.26**	**80.51**	**81.16**
ExpertRAG-Filter + Expert-CoT		$\Delta_{\text{Acc/F1}} = +3.44/+2.59$			
ExpertRAG	FTR	**80.99**	**80.99**	79.45	80.16
ExpertRAG	RTF	80.95	80.96	**79.47**	**80.22**

Table 5: End-to-end RAG Performance and Ablation Study on ExpertRAG and Expert-CoT.

a small number of examples yields substantial gain over the zero-shot baseline. However, adding examples beyond a certain point leads to diminishing or no improvement.

Baseline LLMs underperform on easier questions. As shown in Figure 4a, across all models, we observe that performance is lowest for the EMR certification level (the most basic tier in the NREMT exam) and highest for the Paramedic level. This may be due to smaller data size for EMR level and the procedural nature of EMR questions, whereas Paramedic questions aligning more closely with the medical content seen during LLM pretraining.

LLMs falter on the core NREMT domains. Figure 4b shows that models reliably answer "pharmacology" and "anatomy" questions but struggle in "pediatrics" and core areas such as "trauma","airway", "operations", and "cardiology". One possible reason is subject area complexity. Questions in the former areas often need single-hop, fact-based queries solvable in a zero-shot manner, whereas the latter demand multi-hop reasoning and richer EMS knowledge.

Expertise Classification Performance

The performance of our Filter vs. LLM baselines (0-shot, 4-shot, and CoT) for expertise classification are shown in Table 4. Since subject area classification is a *multi-label* task and certification classification is a *multi-class* task, we report micro f1-score (miF) and macro f1-score (maF). Results show our Filter trained with LoRA with two classification heads significantly outperforms the baseline LLMs.

Model	Description	EMR				EMT				AEMT				Paramedic			
		Pass	Score	Acc	T	Pass	Score	Acc	T	Pass	Score	Acc	T	Pass	Score	Acc	T
Qwen3-4B	0-shot	✗	809	64.18	33	✗	940	74.07	33	✗	940	71.64	33	✗	940	71.72	35
	Expert-CoT	✗	940	72.42	59	✗	940	76.73	73	✓	1179	80.41	76	✗	940	76.03	87
ExpertRAG-4B	FTR+Expert-CoT	✓	1218	84.21	66	✗	940	78.76	61	✗	940	77.31	69	✗	940	79.67	74
	RTF+Expert-CoT	✗	940	76.47	59	✓	1185	81.30	93	✓	1190	83.53	67	✗	940	77.61	86
Qwen3-32B	0-shot	✓	1207	82.65	22	✓	1140	81.63	23	✓	1280	85.92	26	✓	1163	81.58	29
	Expert-CoT	✓	1261	86.27	50	✓	1255	86.96	52	✓	1310	89.11	61	✓	**1292**	**89.01**	57
ExpertRAG-32B	FTR+Expert-CoT	✓	**1350**	**92.22**	75	✓	1292	89.01	76	✓	1215	84.60	86	✓	1228	83.93	125
	RTF+Expert-CoT	✓	**1350**	**92.22**	75	✓	**1328**	**92.32**	82	✓	**1356**	**92.31**	82	✓	1276	88.04	99

Table 6: Pass (✓) or Fail (✗) Summary of Models by Simulation Certification Test. T: Overall Time (min).

Expert-CoT Evaluation

To assess how domain expertise, injected via Expert-CoT, influences reasoning (**RQ2**), we compare Expert-CoT with different prompting strategies under multiple LLMs. As shown in Table 3, **Expert-CoT help guide LLM reasoning.** Integrating domain expert knowledge via CoT-Expert guides reasoning towards appropriate context and consistently boosts CoT prompting performance by up to 2.05% across models. Also, using the predicted expertise attributes (Filter) vs. the ground-truth attribute annotations (GT) yields comparable performance for Expert-CoT, demonstrating the Filter's strong performance, as also shown in Table 4.

Ablation Study on Expert-CoT and ExpertRAG

We further evaluate the effect of injecting expertise into CoT and RAG (**RQ2**) using an ablation study with six configurations, as shown in Table 5: (1) *No-RAG*, (2) *RAG+CoT* with a standard global retriever on EMS PR and KB, (3) *RAG+Expert-CoT*, (4) *ExpertRAG-GT+CoT*, (5) *ExpertRAG-GT+Expert-CoT*, and (6) *ExpertRAG-Filter+Expert-CoT*. An ablation study on certification and subject area is presented in Appendix A.8. The best configuration outperforms the baseline by 4.59 / 3.69 points in Acc / F1 (See error analysis in Appendix A.9).

Effect of PR and KB. (2) vs. (1) isolates the gain of adding PR and KB as retrieval documents with a standard global retriever. Results show KB brings more improvement than PR, and combing both yields the best performance.

Effect of Expert-CoT. (3) vs. (2) (and (5) vs. (4)) ablates the additional gain from Expert-CoT on RAGs, showing that expertise-aware reasoning is better than standard reasoning.

Effect of Expert-RAG. (4) vs. (2) (and (5) vs. (3)) ablate the impact of our Expert-RAG (FTR/RTF) compared to standard RAG with a global retriever. With ground-truth subject area and certification level, ExpertRAG consistently outperforms SOTA RAG baselines, highlighting the value of expertise-guided retrievers. Both FTR and RTF outperform global retrieval, with RTF achieving better performance.

Effect of Filter. (6) vs. (5) measures the effect of using the Filter's predicted expertise vs. ground-truth expertise annotations. There is a small performance drop, but ExpertRAG-Filter still outperforms the best baseline.

NREMT Computer Adaptive Simulation Tests

To investigate whether our best models can be certified in NREMT exam (**RQ3**), we subscribed to MedicTests (MedicTests 2025), the NREMT Computer Adaptive Simulation Test. The simulation exam consists of 80-150 adaptively selected questions and must be completed within 2.5 hours. The NREMT cognitive exam is scored on a 100–1500 scale, with 950 as the passing threshold. We evaluated our Expert-CoT and ExpertRAG models, along with 0-shot baselines. The expertise-augmented models used the trained Filter to predict the subject area and certification. The Pass/Fail outcomes are shown in Table 6. All models completed the exam within the allotted time, though expertise-augmented models took much longer. These models consistently achieved higher test scores and significantly improved accuracy relative to baseline LLMs. However, performance varied with model size. 4B models failed at one or more certification levels, whereas 32B models passed all four. ExpertRAG-32B with the RTF retrieval strategy achieved the highest overall score across certifications. Notably, although the smaller LLM did not pass the test, it benefited the most from expertise augmentation, by showing the largest accuracy gains and achieving scores near or above the passing threshold.

Conclusion

This paper presents a domain expertise-aware LLM framework for medical multiple-choice question answering that infers and incorporates expertise to guide LLM reasoning and RAG retrieval. We introduce **EMSQA**, the first large-scale labeled MCQA dataset for EMS with subject area and certification-level annotations, along with curated EMS knowledge bases. We propose **Expert-CoT**, which guides LLM reasoning by injecting expertise attributes into prompts, and **ExpertRAG**, which retrieves expertise-specific knowledge for augmented generation. Experiments show that our expertise-aware prompting and RAG strategies significantly improve performance over baselines. Importantly, the expertise-augmented LLMs pass the NREMT simulation tests across all EMS certification levels. EMSQA provides a new benchmark for MCQA research in medical domain, and our proposed expertise-aware LLM framework can be applied to other medical MCQA datasets with similar or other expertise attributes.

Ethics Statement

All models studied in this work are research prototypes and not approved medical devices. They must not be used as the sole basis for diagnosis or treatment decisions. Outputs should serve only as a reference for licensed healthcare professionals, who remain fully responsible for clinical judgment and patient care. The models may generate incorrect, incomplete, or biased recommendations and may not reflect up-to-date guidelines. All experiments were conducted in simulation, with no model outputs used to influence real-world patient care. All private data were kept confidential.

Acknowledgments

This work was supported by the award 70NANB21H029 from the U.S. Department of Commerce, National Institute of Standards and Technology (NIST), and a research grant from the Commonwealth Cyber Initiative (CCI).

References

Abd-Alrazaq, A.; AlSaad, R.; Alhuwail, D.; Ahmed, A.; Healy, P. M.; Latifi, S.; Aziz, S.; Damseh, R.; Alrazak, S. A.; Sheikh, J.; et al. 2023. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Medical Education*, 9(1): e48291.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

American Red Cross. 2025. Emergency Medical Response (Red Cross PDF). https://www.redcross.org/content/dam/redcross/training-services/course-fact-sheets/EMR-Textbook-2017-LoRes-111017.pdf. Accessed: 2025-07-14.

Ankit Pal, M. S. 2024. OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Ben Abacha, A.; and Demner-Fushman, D. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20: 1–23.

Bluefield Rescue. 2025. EMT Basic Exam, 5th Edition. http://www.bluefieldrescue.org/EMTBasicExam5thEdition.pdf. Accessed: 2025-07-14.

Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1): D267–D270.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cai, Y.; Wang, L.; Wang, Y.; de Melo, G.; Zhang, Y.; Wang, Y.; and He, L. 2024. Medbench: A large-scale chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17709–17717.

CareerEmployer. 2025. AEMT Practice Test—CareerEmployer. https://careeremployer.com/test-prep/practice-tests/aemt-practice-test/. Accessed: 2025-07-14.

Carrie Davis. 2025. Carrie Davis YouTube Channel. https://www.youtube.com/@CarrieDavis. Accessed: 2025-07-14.

Chan, C.-M.; Xu, C.; Yuan, R.; Luo, H.; Xue, W.; Guo, Y.; and Fu, J. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Dawson, D. E. 2006. National emergency medical services information system (NEMSIS). *Prehospital Emergency Care*, 10(3): 314–316.

Delaware Health and Social Services. 2025. Paramedic Medication Manual (Delaware). https://www.dhss.delaware.gov/dph/ems/files/PHARMACOLOGYMANUAL2024.pdf. Accessed: 2025-07-14.

DocDrop / Arthur Hsieh. 2025. EMT Exam for Dummies (with Online Practice). https://docdrop.org/static/drop-pdf/EMT-Exam-For-Dummies-with-Online-Practice---Arthur-Hsieh-Qvn0s.pdf. Accessed: 2025-07-14.

Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R. O.; and Larson, J. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

EMRA. 2025. EMS Essentials: A Resident's Guide to Prehospital Care. https://www.emra.org/siteassets/emra/publications/books/emra-ems-essentials.pdf. Accessed: 2025-07-14.

EMT-Prep. 2025. EMT-Prep App. https://app.emtprep.com/. Accessed: 2025-07-14.

Es, S.; James, J.; Anke, L. E.; and Schockaert, S. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158.

Fenniak, M.; Stamy, M.; pubpub zz; Thoma, M.; Peveler, M.; exiledkingcc; and PyPDF2 Contributors. 2022. The PyPDF2 library. https://pypi.org/project/PyPDF2/.

Firearms Training Los Angeles CA. 2025. Emergency Medical Responder: Your First Response in Emergency Care. https://firearmstraininglosangelesca.com/wp-content/uploads/2019/07/AAOS_First_Responder_eBook1.pdf. Accessed: 2025-07-14.

Ge, X.; Satpathy, A.; Williams, R.; Stankovic, J.; and Alemzadeh, H. 2024. DKEC: Domain Knowledge Enhanced Multi-Label Classification for Diagnosis Prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 12798–12813.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jeong, M.; Sohn, J.; Sung, M.; and Kang, J. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1): i119–i129.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Jin, Q.; Kim, W.; Chen, Q.; Comeau, D. C.; Yeganova, L.; Wilbur, W. J.; and Lu, Z. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11): btad651.

Jones & Bartlett Learning. 2025. JB Learning EMS Slides. https://www.jblearning.com/. Accessed: 2025-07-14.

Khashabi, D.; Chaturvedi, S.; Roth, M.; Upadhyay, S.; and Roth, D. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 252–262.

Krithara, A.; Nentidis, A.; Bougiatiotis, K.; and Paliouras, G. 2023. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Scientific Data*, 10(1): 170.

Kung, T. H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2): e0000198.

LearningExpress Hub. 2025. LearningExpress Hub EMS Prep. https://www.learningexpresshub.com/ProductEngine/LELIndex.html#/center/learningexpresslibrary/career-center/home/prepare-for-emergency-medical-services-and-firefighting-exams. Accessed: 2025-07-14.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Li, Z.; Chen, X.; Yu, H.; Lin, H.; Lu, Y.; Tang, Q.; Huang, F.; Han, X.; Sun, L.; and Li, Y. 2025a. StructRAG: Boosting Knowledge Intensive Reasoning of LLMs via Inference-time Hybrid Information Structurization. In *The Thirteenth International Conference on Learning Representations*.

Li, Z.; Luo, X.; Ge, X.; Zhou, L.; Lin, X.; and Liu, Y. 2025b. MMSense: Adapting Vision-based Foundation Model for Multi-task Multi-modal Wireless Sensing. arXiv:2511.12305.

Liu, S.; Johns, E.; and Davison, A. J. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1871–1880.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, Y.; Zhao, X.; and Wang, J. 2024. ClinicalRAG: Enhancing Clinical Decision Support through Heterogeneous Knowledge Retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, 64–68.

Luo, X.; Jha, S. M. N.; Sinha, A.; Li, Z.; and Liu, Y. 2025. ALPHA: LLM-Enabled Active Learning for Human-Free Network Anomaly Detection. *arXiv preprint arXiv:2509.05936*.

MedicTests. 2025. Exact replica of the NREMT Simulator. https://medictests.com/. Accessed: 2025-07-18.

MediPro First Aid. 2025. Professional Responder Cheat Sheet. https://www.mediprofirstaid.net/free-downloads/Professional_Responder_Cheat_Sheet.pdf. Accessed: 2025-07-14.

MinniQuiz. 2025. First Responder Practice Sets. Available at: https://minniquiz.com/First_Responder/; https://minniquiz.com/First_Responder1/; https://minniquiz.com/First_Responder2/; https://minniquiz.com/First_Responder_Paramedic/. Accessed: 2025-07-14.

Mometrix Academy. 2025. Mometrix EMT & Paramedic Practice. Available at: https://www.mometrix.com/academy/advanced-emt; https://www.mometrix.com/academy/emt-practice-test; https://www.mometrix.com/academy/emt-exam-paramedics. Accessed: 2025-07-14.

Montgomery County MD. 2025. Montgomery County NREMT Resource. https://www.montgomerycountymd.gov/mcfrs-psta/Resources/Files/Instructor/NR_EMT_Test.htm. Accessed: 2025-07-14.

My CPR Certification Online. 2025. Practice Tests: EMR, EMT, and Paramedic. Available at: https://www.mycprcertificationonline.com/practice-tests/emr; https://www.mycprcertificationonline.com/practice-tests/emt; https://www.mycprcertificationonline.com/practice-tests/paramedics. Accessed: 2025-07-14.

NREMT. 2001–2025. National Registry of Emergency Medical Technicians. https://www.nremt.org/. Accessed: 2025-07-18.

NREMT Practice Test. 2025. NREMT Practice Test. https://nremtpracticetest.com/. Accessed: 2025-07-14.

ODEMSA. 2025. ODEMSA Regional EMS Documents. https://odemsa.net/regional-documents/. Accessed: 2025-07-14.

Ou, J.; Huang, T.; Zhao, Y.; Yu, Z.; Lu, P.; and Ying, R. 2025. Experience Retrieval-Augmentation with Electronic Health Records Enables Accurate Discharge QA. *arXiv preprint arXiv:2503.17933*.

Oudenhove, L. V. 2024. Enhancing RAG Performance with Metadata: The Power of Self-Query Retrievers. https://medium.com/@lorevanoudenhove/enhancing-rag-performance-with-metadata-the-power-of-self-query-retrievers-e29d4eecdb73. Accessed: 2025-11-14.

Pal, A.; Minervini, P.; Motzfeldt, A. G.; and Alex, B. 2024. Open Medical LLM Leaderboard. https://github.com/openlifesciencesai/open_medical_llm_leaderboard.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260. PMLR.

Pampari, A.; Raghavan, P.; Liang, J.; and Peng, J. 2018. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2357–2368. Brussels, Belgium: Association for Computational Linguistics.

PocketPrep. 2025. PocketPrep EMT/Paramedic. https://www.pocketprep.com/. Accessed: 2025-07-14.

PracticeTestGeeks. 2025. PracticeTestGeeks EMR. https://practicetestgeeks.com/emr-practice-test/. Accessed: 2025-07-14.

Preum, S. M.; Shu, S.; Alemzadeh, H.; and Stankovic, J. A. 2020. Emscontext: EMS protocol-driven concept extraction for cognitive assistance in emergency response. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13350–13355.

Quizizz. 2025. Quizizz. https://quizizz.com. Accessed: 2025-07-14.

Quizlet. 2025. Quizlet Question Banks for EMR, EMT, AEMT, and Paramedic. Available at: https://quizlet.com/search?query=emr-practice-test&type=questionBanks; https://quizlet.com/search?query=emt-practice-test&type=questionBanks; https://quizlet.com/search?query=aemt-practice-test&type=questionBanks; https://quizlet.com/search?query=paramedics-practice-test&type=questionBanks. Accessed: 2025-07-14.

Rhode Island Department of Health. 2025. EMS Pharmacology Reference Guide. https://health.ri.gov/sites/g/files/xkgbur1006/files/publications/guides/EMSPharmacologyReference.pdf. Accessed: 2025-07-14.

SmartMedic. 2025. SmartMedic EMT Practice. https://smartmedic.com/index.php. Accessed: 2025-07-14.

Sohn, J.; Park, Y.; Yoon, C.; Park, S.; Hwang, H.; Sung, M.; Kim, H.; and Kang, J. 2024. Rationale-Guided Retrieval Augmented Generation for Medical Question Answering. *arXiv preprint arXiv:2411.00300*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, Q. 2025. Qwen3 Technical Report. arXiv:2505.09388.

The EMS Professor. 2025. The EMS Professor YouTube Channel. https://www.youtube.com/@theemsprofessor. Accessed: 2025-07-14.

Union Test Prep. 2025. Union Test Prep EMT Practice. https://uniontestprep.com/emt-test/practice-test. Accessed: 2025-07-14.

Weerasinghe, K.; Janapati, S.; Ge, X.; Kim, S.; Iyer, S.; Stankovic, J. A.; and Alemzadeh, H. 2024. Real-Time Multimodal Cognitive Assistant for Emergency Medical Services. In *2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI)*, 85–96. IEEE.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wiley. 2025. Emergency Medical Services: Clinical Practice and Systems Oversight. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118990810. Accessed: 2025-07-14.

Xiong, G.; Jin, Q.; Lu, Z.; and Zhang, A. 2024a. Benchmarking Retrieval-Augmented Generation for Medicine. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 6233–6251. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Xiong, G.; Jin, Q.; Wang, X.; Zhang, M.; Lu, Z.; and Zhang, A. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, 199–214. World Scientific.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623.

Appendix

A.1 Introduction

This is the technical appendix for the paper "Expert-Guided Prompting and Retrieval-Augmented Generation for Emergency Medical Service Question Answering".

A.2 EMSQA Data

A.2.1 EMSQA Data Collection

In Table 7, we present detailed information on 17 public and private EMSQA resources, including each source's name and URL, number of questions, certification levels, and the availability of explanations.

A.2.2 EMSQA Data Statistics

In Table 8, we show the detailed EMSQA dataset statistics. The dataset includes a total of 18,602 and 5,669 practice questions from public and private sources, respectively. We use the private questions exclusively for testing and split the public questions into train, validation, and test sets of 13,021, 1,860, 3,721 questions, with average token lengths of 18.27, 19.12, 18.99, respectively.

A.3 Knowledge Base Preprocessing & Statistics

A.3.1 Prompt for Organizing Chapter Text

As shown in Figure 5, we present the prompt that instructs the GPT-4o to divide each chapter's raw text into clearly labeled sections.

PROMPT:
Well formated the unstructured text by the folllowing rules:
1. Fix awkward or broken line breaks within sentences or paragraphs.
2. Separate paragraphs with a single blank line.
3. Remove figure captions and references (e.g., "Figure 8–1", "see Figure…").
4. Remove page numbers (e.g., "239").
5. Remove attribution or copyright lines, such as: defined copyrights
6. **Do not rephrase, reword, rewrite, or summarize; do not add any other words in your response**.
7. If the unstructured text contains subtitles, treat each subtitle as a key and store its corresponding paragraph(s) as the value. Return only the resulting JSON.
Here is the unstructured text to format: **RAW TEXT**

Figure 5: Prompt for organizing chapter text

A.3.2 Prompt for EMS Concept Extraction

In main paper **Table 3: Semantic and syntactic evaluation of QA overlap vs. KB/PR**, we present the concept overlap in Syntactic metrics, we present the prompt we used for EMS concept extraction in Figure 6.

A.3.3 UMLS Concept Normalization

We applied UMLS API (Bodenreider 2004) to normalize the extracted EMS concepts and retained only terms whose semantic types fell into one of the following categories: Sign or Symptom (Sosy), Finding (Fndg), Laboratory or Test Result (Lbtr), Clinical Attribute (Clna), Quantitative Concept (Qnco), Qualitative Concept (Qlco), Disease or Syndrome (Dsyn), Mental or Behavioral Dysfunction (Mobd), Pathologic Function (Patf), Neoplastic Process (Neop), Congenital Abnormality (Cgab), Anatomical Abnormality (Anab), Injury or Poisoning (Inpo), Cell or Molecular Dysfunction (Celf), Body Part/Organ/Organ Component (Bpoc), Body Location or Region (Bodl), Body Space or Junction (Bsoj), Body System (Bodsys), Therapeutic or Preventive Procedure (Topp), Diagnostic Procedure (Diap), Laboratory Procedure (Lbpr), Imaging Procedure (Impr), Health Care Activity (Hlca), Clinical Drug (Clnd), Pharmacologic Substance (Phsu), Antibiotic (Antb), Vitamin (Vita), Organic Chemical (Orch), Amino Acid/Peptide/Protein (Aapp), Biologically Active Substance (Bacs), Hazardous or Poisonous Substance (Hops), Steroid (Strd), Hormone (Horm), Medical Device (Medd), Temporal Concept (Tmco), Spatial Concept (Spco), and Functional Concept (Fngp).

A.3.4 Knowledge and EMSQA Overlap Statistics

The detailed statistics such as the total number of concepts in EMSQA and KB, overlapped concepts (KB $\cap$ QA), union concepts (KB $\cup$ QA), KB unique concepts (KB $\setminus$ QA) are shown in Table 9. The syntactic (hit rate) reported in paper is calculated by KB $\cap$ QA / QA.

A.3.5 Knowledge Base Collection Details

Table 10 shows the detailed information of every knowledge source we collected to construct external EMS knowledge bases. Specifically, we list the resource type, title, URL, certification levels, and vocabulary of each source. In total, there are 2,545,192 tokens in our external knowledge bases, and the unique vocabulary size is 34,110.

Source	# QAs	Certification Level(s)	Explanation
Public			
(Bluefield Rescue 2025)	955	EMT	✓
(NREMT Practice Test 2025)	240	EMT	✓
(SmartMedic 2025)	537	EMT	✗
(Union Test Prep 2025)	150	EMT	✓
(Montgomery County MD 2025)	130	EMT	✓
(Mometrix Academy 2025)	138	EMT, AEMT, Paramedic	✓
(PracticeTestGeeks 2025)	99	EMR	✓
(My CPR Certification Online 2025)	74	EMR, EMT, AEMT	✗
(Quizizz 2025)	1,554	EMR, EMT, AEMT, Paramedic	✗
(MinniQuiz 2025)	4,444	EMR, EMT, Paramedic	✗
(Quizlet 2025)	9,084	EMR, EMT, AEMT, Paramedic, NA	✗
(CareerEmployer 2025)	100	AEMT	✓
(DocDrop / Arthur Hsieh 2025)	456	EMT	✓
(LearningExpress Hub 2025)	1,336	Paramedic, EMT	✓
(PocketPrep 2025)	47	EMR, EMT, AEMT, Paramedic	✓
Private			
(EMT-Prep 2025)	4,843	EMR, EMT, AEMT, Paramedic, Critical Care	✓
(Jones & Bartlett Learning 2025)	1,285	EMT	✓

Table 7: Summary of NREMT practice question sources in EMSQA.

Data	Split	#Explanations	#Choices (avg/max)	#Answers (avg/max)	Question Tokens (avg/max)	Choice Tokens (avg/max)	Tokens	Vocab
Public	Train (13,021)	2217	4.01 / 7.00	1.00 / 3.00	18.27 / 218	6.28 / 240	565,303	14,017
	Val (1,860)	383	3.99 / 5.00	1.00 / 3.00	19.12 / 155	6.01 / 44	80,215	6,629
	Test (3,721)	773	4.01 / 6.00	1.00 / 3.00	18.99 / 135	6.10 / 60	161,464	8,913
	Total (18,602)	3132	4.01 / 7.00	1.00 / 3.00	18.50 / 218	6.22 / 240	806,982	16,032
Private	Test (5,669)	5451	4.01 / 6.00	1.06 / 4.00	30.44 / 355	5.46 / 47	296,673	10,637

Table 8: Statistics by split for Public and Private EMSQA

Data	Comparison	QA	KB	PR	KB ∩ QA	PR ∩ QA	KB \ QA	PR \ QA	KB ∪ QA	PR ∪ QA
Public	Vocab	15,892	33,965	6,773	13,183	3,359	20,782	3,414	36,674	19,306
	Cpts (w/o norm)	25,594	61,003	9,383	10,661	2,269	50,342	7,114	75,936	32,708
	Cpts (w norm)	17,262	34,627	8,088	10,927	2,637	23,700	5,451	40,962	22,713
Private	Vocab	10,496	33,965	6,773	9,540	2,966	24,425	3,807	34,921	14,303
	Cpts (w/o norm)	11,621	61,003	9,383	6,180	1,669	54,823	7,714	66,444	19,335
	Cpts (w norm)	8,689	34,627	8,088	6,299	1,969	28,328	6,119	37,017	14,808

Table 9: Detailed overlap statistics between QA and KB for public and private data.

A.4 NEMSIS Patient Care Report Preprocessing & Statistics

Access to NEMSIS 2021 Public-Release Research Dataset requires submitting a request on the NEMSIS website[1].

A.4.1 NEMSIS Patient Record Preprocessing

To concatenate information for each patient record in the NEMSIS dataset, we used the following fields in different tables and concatenated them by the shared primary key. The keys in our patient records include "Date - Time of Symptom Onset", "Chief Complaints", "Possible Injury", "Cause of Injury", "Chief Complaint Anatomic Location", "Chief Complaint Organ System", "Gender", "Age", "Age Unit", "Level of Responsiveness (AVPU)", "Primary Symptoms", "Other Associated Symptoms", "Primary Impressions", "Secondary Impressions", "Protocol Age Category", "Protocols", "Date - Time Vital Signs Taken", "ECG Type", "SBP (Systolic Blood Pressure)", "Heart Rate", "Respiratory Rate", "Pulse Oximetry", "Blood Glucose Level", "End

[1] https://nemsis.org/using-ems-data/request-research-data/

Resource Type and Title	Level	# Vocab
Transcripts		
Emergency Care and Transportation of the Sick and Injured Advantage Package (Carrie Davis 2025)	EMT	23,836
Nancy Caroline's Emergency Care in the Streets (Carrie Davis 2025)	Paramedic	
AAOS Advanced Emergency Medical Technician (AEMT) 4th Ed (The EMS Professor 2025)	AEMT	
AAOS Critical Care Transport Paramedic (The EMS Professor 2025)	Paramedic	
Textbooks		
EMT Exam for Dummies (DocDrop / Arthur Hsieh 2025)	EMT	2,025
Emergency Medical Services: Clinical Practice and Systems Oversight (Wiley 2025)	—	18,799
EMS Essentials: A Resident's Guide to Prehospital Care(EMRA 2025)	—	4,123
Emergency Medical Response - RedCross (American Red Cross 2025)	EMR	10,096
Emergency Medical Responder: Your First Response in Emergency Care (Firearms Training Los Angeles CA 2025)	EMR	7,794
Guidelines		
Professional Responder Cheat Sheet (MediPro First Aid 2025)	EMR	1,569
EMS Pharmacology Reference Guide (Rhode Island Department of Health 2025)	—	3,072
Paramedic Medication Manual (Delaware Health and Social Services 2025)	Paramedic	2,011
ODEMSA Regional EMS Documents (ODEMSA 2025)	—	6,340
Slides		
Emergency Care and Transportation of the Sick and Injured Advantage Package (Jones & Bartlett Learning 2025)	EMT	8,597
Flashcards & Study Guides		
EMS Study Guides (EMT-Prep 2025)	EMR, EMT, AEMT, Paramedic	8,373

Table 10: Knowledge base collection details

Tidal Carbon Dioxide (ETCO2)", "Glasgow Coma Score-Eye", "Glasgow Coma Score-Verbal", "Glasgow Coma Score-Motor", "Pain Scale Score", "Stroke Scale Score", "Stroke Scale Type", "Reperfusion Checklist", "Date - Time Medication Administered", "Medication Administered Prior to this Unit's EMS Care", "Medication Given", "Medication Dosage", "Medication Dosage Units", "Response to Medication", "Role - Type of Person Administering Medication", "Date - Time Procedure Performed", "Procedure", "Number of Procedure Attempts", "Response to Procedure", "Role - Type of Person Performing the Procedure", "Alcohol - Drug Use Indicators", "Barriers to Patient Care", "Cardiac Arrest", "Date - Time of Cardiac Arrest", "First Monitored Arrest Rhythm of the Patient", "Cardiac Arrest Etiology", "Type of CPR Provided", "Cardiac Rhythm on Arrival at Destination", "Reason CPR - Resuscitation Discontinued", "Incident - Patient Disposition", "EMS Transport Method", "Transport Mode from Scene", "Initial Patient Acuity", "Final Patient Acuity".

We further classify patient records to seven subject areas by the field protocol EMS. Specifically,

"Assessment" includes "General-Universal Patient Care/ Initial Patient Contact", "General-Individualized Patient Protocol"

Medical & OB includes "Medical-Seizure", "Medical-Nausea/Vomiting", "Medical-Influenza-Like Illness/ Upper Respiratory Infection", "Medical-Abdominal Pain", "Medical-Altered Mental Status", "OB/GYN-Pregnancy Related Emergencies", "Medical-Supraventricular Tachycardia (Including Atrial Fibrillation)", "Medical-Cardiac Chest Pain", "Medical-Tachycardia", "Medical-Syncope", "Medical-Stroke/TIA", "Medical-Respiratory Distress/Asthma/COPD/Reactive Airway", "Medical-Respiratory Distress-Bronchiolitis", "Medical-Diarrhea", "Medical-Hypoglycemia/Diabetic Emergency", "Medical-Hyperglycemia", "Medical-Allergic Reaction/Anaphylaxis", "Medical-Hypertension", "Medical-Bradycardia", "Medical-Pulmonary Edema/CHF", "Medical-Hypotension/Shock (Non-Trauma)", "Medical-Opioid Poisoning/Overdose", "Medical-ST-Elevation Myocardial Infarction (STEMI)", "Medical-Ventricular Tachycardia (With Pulse)", "Medical-Stimulant Poisoning/Overdose", "Medical-Respiratory Distress-Croup", "Medical-Adrenal Insufficiency", "Medical-Beta Blocker Poisoning/Overdose", "Medical-Calcium Channel Blocker Poisoning/Overdose", "OB/GYN-Gynecologic Emergencies", "OB/GYN-Childbirth/Labor/Delivery", "OB/GYN-Eclampsia", "OB/GYN-Post-partum Hem-

PROMPT:
Extract all EMS concepts from the following text.
Here are some examples of EMS concepts: gelastic epilepsy, visual seizure, pallor, pale color, aox4, pare down, cut down
Return all the extracted EMS concepts as a lowercase letter in strict JSON format, like: ["fever", "cardiac arrest"]

Here is one example:
Text: Early symptoms include cough, wheezing, shortness of breath. Lung transplantation is an option. Response: Let's think step by step,

Step1: label the tokens one by one "EMS concept", or "none".
-Early: none
-symptoms: none
-include: none
-cough: EMS concept
-wheezing: EMS concept
-shortness: EMS concept
-of: none
-breath: EMS concept
-Lung: EMS concept
-transplantation: EMS concept
-is: none
-an: none
-option: none

Step2: RefineEMS concept from Step 1 by following criteria,
1.concatenate EMS concept spans
2.remove extra irrelevant words in EMS concept
-cough: EMS concept
-wheezing: EMS concept
-shortness of breath: EMS concept
-Lung transplantation: EMS concept

Step3: Return the your result: ["cough", wheezing", "shortness of breath", "Lung transplantation"]
Now is the real text: **RAW TEXT**

Figure 6: Prompt for EMS Concept Extraction

orrhage", "General-Overdose/Poisoning/Toxic Ingestion", "General-Fever", "General-Epistaxis", "General-Back Pain", "General-Pain Control", "Environmental-Altitude Sickness", "Environmental-Cold Exposure", "Environmental-Hypothermia", "Environmental-Heat Stroke/Hyperthermia", "Environmental-Heat Exposure/Exhaustion"

"Airway" includes "Airway", "Airway-Failed", "Airway-Sedation Assisted (Non-Paralytic)", "Airway-Rapid Sequence Induction (RSI-Paralytic)", "Airway-Obstruction/Foreign Body",

"EMS Operations" includes "Exposure-Airway/Inhalation Irritants", "General-Neglect or Abuse Suspected", "Exposure-Explosive/ Blast Injury", "General-IV Access", "General-Refusal of Care", "General-Behavioral/Patient Restraint", "General-Interfacility Transfers", "General-Spinal Immobilization/Clearance", "General-Medical Device Malfunction", "General-Law Enforcement - Assist with Law Enforcement Activity", "General-Extended Care Guidelines", "General-Exception Protocol", "General-Indwelling Medical Devices/Equipment", "General-Community Paramedicine / Mobile Integrated Healthcare", "General-Law Enforcement - Blood for Legal Purposes", "General-Dental Problems", "Exposure-Biological/Infectious", "Exposure-Carbon Monoxide"", "Exposure-Chemicals to Eye", "Exposure-Smoke Inhalation", "Exposure-Cyanide", "Exposure-Radiologic Agents", "Exposure-Blistering Agents", "Exposure-Nerve Agents"

"Cardiovascular" includes "General-Cardiac Arrest", "Cardiac Arrest-Asystole", "Cardiac Arrest-Determination of Death / Withholding Resuscitative Efforts", "Cardiac Arrest-Pulseless Electrical Activity", "Cardiac Arrest-Do Not Resuscitate", "Cardiac Arrest-Ventricular Fibrillation/ Pulseless Ventricular Tachycardia", "Cardiac Arrest-Post Resuscitation Care", "Cardiac Arrest-Special Resuscitation Orders", "Cardiac Arrest-Hypothermia-Therapeutic"

"Pediatrics" includes "Medical-Apparent Life Threatening Event (ALTE)", "Medical-Newborn/ Neonatal Resuscitation"

"Trauma" includes "Environmental-Frostbite/Cold Injury", "Injury-Head", "Injury-Extremity", "Injury-Burns-Thermal", "Injury-General Trauma Management", "Injury-Multisystem", "Injury-Eye", "Injury-Mass/Multiple Casualties", "Injury-Amputation", "Injury-Spinal Cord", "Injury-Conducted Electrical Weapon (e.g., Taser)", "Injury-Bleeding/ Hemorrhage Control", "Injury-Bites and Envenomations-Land", "Injury-Facial Trauma", "Injury-Cardiac Arrest", "Injury-Crush Syndrome", "Injury-Thoracic", "Injury-Drowning/Near Drowning", "Injury-Diving Emergencies", "Injury-Electrical Injuries", "Injury-Topical Chemical Burn", "Injury-Impaled Object", "Injury-Bites and Envenomations-Marine", "Injury-Lightning/Lightning Strike", "Injury-SCUBA Injury/Accidents".

NEMSIS Patient Records	
Total cases	4,003,430
Total tokens	1,247,811,207
Average tokens/case	311.7
Min tokens	202
Max tokens	824
Vocabulary	6,838

Table 11: Statistics for the NEMSIS patient care reports

A.4.2 NEMSIS Patient Record Statistics

The detailed statistics of processed NEMSIS data is shown at Table 11. There are in total 4,003,430 patient records with

Dimension	Scale	Expert Rating
Clinical Accuracy	0-100	100
Difficulty Level	1-5	3.95
Subject Area Accuracy	1-5	4.85
Knowledge Relevance	1-5	4.39

Table 12: Human Expert evaluation on EMSQA & KB.

average 311.7 tokens per case. The total number vocabulary is 6,838.

We also show the statistics such as the total number of concepts in EMSQA and PR, overlapped concepts (PR $\cap$ QA), union concepts (PR $\cup$ QA), PR unique concepts (PR $\setminus$ QA) are shown in Table 9. The syntactic (hit rate) reported in paper is calculated by PR $\cap$ QA / QA.

A.5 Human Evaluation of EMSQA and Knowledge Base

To validate the quality and clinical utility of our EMSQA benchmark and its accompanying curated knowledge bases, we conducted an expert review on a randomly sampled subset of 100 questions. Sampling was stratified to ensure balanced coverage across certification levels—25 questions each for EMR, EMT, AEMT, and Paramedic—and across five major medical subject areas—20 questions each for Airway, Trauma, Cardiology, Medical, and Operations—while also drawing uniformly from all source links.

We asked one EMT-certified professional to rate each question–answer pair on four dimensions:

Clinical Accuracy: Verify that the provided "correct" answer is clinically sound. *Rating:* Yes / No

Difficulty Level: Judge whether the question's difficulty matches the expected level for its certification. *Rating:* 1 (Strongly Disagree) to 5 (Strongly Agree)

Subject Area Accuracy: Confirm that the assigned medical subject area appropriately reflects the question's content. *Rating:* 1 (Strongly Disagree) to 5 (Strongly Agree)

Knowledge Relevance: Assess whether the assembled knowledge base sufficiently supports answering the question. *Rating:* 1 (Strongly Disagree) to 5 (Strongly Agree)

Table 12 shows average rating for four dimensions. The results indicate that: (1) provided answers are clinically accurate; (2) the EMT expert largely agrees with the assigned certification levels and subject areas; and (3) the curated knowledge base is helpful for answering the questions.

A.6 Retrieval Strategy Performance

In the paper we present three different retrieval strategies (**Global**, **Filter then Retrieve (FTR)**, **Retrieve then Filter (RTF)**). To evaluate the retrieval performance, we uniformly sampled 100 questions from the 4 different certification levels (25 questions per certification level) and 10 subject areas ($\sim$ 10 questions per subject area). For each question, we use MedCPT to retrieve 32 documents from Knowledge Bases.

Since we do not have the ground-truth question-document pairs, we use LLM-as-a-judge (Zheng et al. 2023) to evaluate the **"Relevance"** (Es et al. 2024; Asai et al. 2023) and **"Supportiveness"** (Asai et al. 2023) of the 32 retrieved documents for every question. Relevance judges if each retrieved chunk is pertinent to the question, while supportiveness judges if this retrieved chunk helps answer the question.

As shown in Figure 7, we present the prompt template we used in LLM-as-a-judge. To evaluate "Relevance" and "Supportiveness" we use four Information Retrieval metrics: Top-k Hit rate ($hit@k$), which measures the fraction of questions for which at least one positive chunk appears in the top-k retrieved; Top-k Precision ($p@k$), which computes the average proportion of positive chunks among the top-k for each question; Mean Reciprocal Rank (MRR), which takes the reciprocal of the rank position of the first positive chunk for each question and then averages these reciprocals—thereby reflecting how early in the ranking the first positive appears; and Mean Average Precision (MAP), which for each question computes the average of precisions at every positive chunk's position and then averages those per-question APs, capturing both the ranking quality and distribution of positives over the entire list. We omit Top-k Recall ($r@k$) since we lack ground-truth documents.

For each question q in sampled questions $\mathcal{Q}$ and retrieved chunk i, we define:

$$r_{q,i}^{\mathrm{rel}} = \begin{cases} 1 & \text{if chunk } i \text{ is Relevant,} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

$$r_{q,i}^{\mathrm{sup}} = \begin{cases} 1 & \text{if chunk } i \text{ is Fully or Partially support,} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Then, for either $\{r_{q,i}^{\mathrm{rel}}\}$ or $\{r_{q,i}^{\mathrm{sup}}\}$:

PROMPT:
Question: {}
Answer: {}
Retrieved Chunk: {}
Rubric
- Relevance
• "Relevant" — The chunk provides information that is useful for answering the question.
• "Irrelevant" — The chunk is off-topic or supplies no helpful information.
- Supportiveness
Evaluate how much of a hypothetical answer could be *entailed* by this chunk alone.
• "Fully" — All necessary information is present; the answer would be completely supported.
• "Partially" — Some but not all required information is present.
• "None" — No information is supported, or the chunk contradicts the answer.

Figure 7: Prompt for LLM-as-a-judge

Strategy	Relevance Hit@k (1 / 5 / 10)	P@k (1 / 5 / 10)	MRR	MAP
Global	44.0 / 69.0 / 78.0	44.0 / 35.0 / 31.7	55.6	39.4
FTR	44.0 / 71.0 / 80.0	44.0 / 35.0 / 32.2	**57.6**	39.6
RTF	**45.0 / 75.0 / 81.0**	**45.0 / 35.0 / 32.5**	57.5	**39.6**
Strategy	**Supportiveness Hit@k (1 / 5 / 10)**	**P@k (1 / 5 / 10)**	**MRR**	**MAP**
Global	25.0 / 52.0 / 67.0	25.0 / 21.0 / 20.3	38.3	27.6
FTR	27.0 / **59.0** / **70.0**	27.0 / **24.0** / 21.2	39.7	**29.2**
RTF	**29.0** / 58.0 / 70.0	**29.0** / 23.0 / **21.6**	**39.8**	28.8

Table 13: Retrieval Relevance and Supportiveness

$$hit@k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}\Big(\sum_{i=1}^{k} r_{q,i} > 0\Big) \tag{9}$$

$$p@k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{k} \sum_{i=1}^{k} r_{q,i} \tag{10}$$

$$\text{MRR} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{\min\{\, i : r_{q,i} = 1\}} \tag{11}$$

$$\text{MAP} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{\sum_{i=1}^{N} r_{q,i}} \sum_{i=1}^{N} r_{q,i} \frac{1}{i} \sum_{j=1}^{i} r_{q,j}\,. \tag{12}$$

Table 13 presents the performance of three retrieval strategies,Global (direct retrieval), Retrieve-then-Filter (RTF) and Filter-then-Retrieve (FTR), in terms of both "Relevance" and "Supportiveness". In terms of "Relevance", RTF and FTR show slightly better performance than Global across all metrics, indicating that the proposed strategies retrieve relevant chunks earlier and more reliably. In terms of "Supportiveness", both RTF and FTR achieve higher Hit@k and P@k and slightly better MRR and MAP than Global, suggesting they more effectively capture fully or partially supportive evidence. Overall, although based on a small study, these results show that the proposed retrieval strategies improve the extraction of both relevant and supportive context from the knowledge base, providing stronger grounding for the LLM's final answer. A more comprehensive evaluation of retrieval strategies is beyond the scope of this paper and left for future work.

A.7 Benchmark LLMs

A.7.1 Zero-shot Performance per Subject Area

As shown in Figure 8, we show the zero-shot LLMs' per subject area performance on private dataset. We only run open-source models on our private dataset. The conclusion remains consistent on the private dataset: **LLMs excel in pharmacology and anatomy but falter on core NREMT domains**. Models answer pharmacology and anatomy items reliably, yet stumble on "trauma", "airway management", "EMS operations", and "medical&OB" questions. One reason may be task complexity: the former are largely single-hop, fact-based queries, whereas the latter demand multi-hop reasoning and richer EMS context.

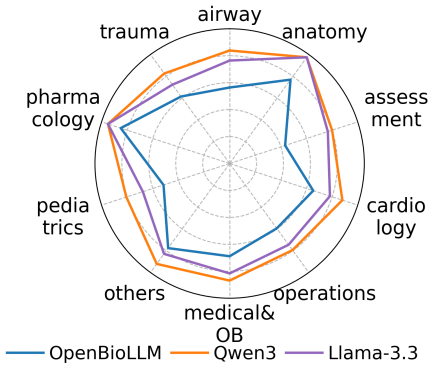

Figure 8: Zero-shot Performance per Subject Area on EMSQA-Private Dataset

A.7.2 Detailed statistics of Benchmarks

Table 14 summarizes the full benchmarking results for each model and prompting strategy, reporting accuracy and F1 across all four EMS certification levels (EMR, EMT, AEMT, Paramedic) as well as overall performance. Below are several key findings:

Closed-source models outperform open-source models. In particular, OpenAI-o3 consistently achieves the highest overall accuracy of 92.39. Among open-source models, Qwen3-32B achieves the best accuracy of 85.70, though a significant gap remains compared to closed-source models.

Few-shot prompting improves accuracy up to a point. We varied the number of in-context exemplars from 0 to 64 and observed that incorporating a small number of examples yields substantial gain over the zero-shot baseline. However, adding examples beyond a certain point leads to diminishing or no improvement.

Expert-CoT helps guide LLM reasoning. Integrating domain expert knowledge via CoT-Expert guides reasoning towards appropriate context and consistently boosts CoT prompting performance by up to 2.05% across models. Also, using the Filter's predicted expertise for Expert-CoT leads to comparable performance to when using ground-truth.

A.8 Ablation Study on Certification Level and Subject Area

To determine which expertise attribute contributes the most to performance, we ablate the two key attributes, "Subject Area" and "Certification", and measure their individual effects on Expert-CoT. We cannot do the ablation study on Expert-RAG since it only uses the subject area for retrieval. The results are presented in Table 15.

Expertise attributes are most effective when used together. Individually, "Subject Area" and "Certification" improve performance in Expert-CoT between 0.74-4.91% in

Model	Prompt	EMSQA—Public									
		EMR		EMT		AEMT		Paramedic		Overall	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Qwen3-32B	0-shot	79.08	79.08	82.98	82.98	83.68	83.68	86.69	86.69	83.55	83.55
	4-shot	80.89	80.89	84.31	84.31	86.05	86.05	86.06	86.06	84.41	84.41
	8-shot	80.20	80.20	84.60	84.60	85.79	85.79	86.14	86.14	84.39	84.39
	16-shot	81.03	81.03	84.23	84.23	85.79	85.79	**87.31**	87.31	84.82	84.82
	32-shot	77.68	77.68	81.13	81.13	81.58	81.58	83.01	83.01	81.13	81.13
	64-shot	80.33	80.33	80.09	80.09	85.26	85.26	85.43	85.43	82.48	82.48
	CoT	80.06	80.06	82.90	82.95	85.26	85.26	86.92	86.92	84.96	84.97
(GT)	Expert-CoT	**82.43**	**82.48**	**85.49**	**85.49**	**87.89**	**87.89**	87.16	87.16	**85.70**	**85.71**
(Filter)	Expert-CoT	81.52	81.51	82.24	83.68	86.37	86.65	86.98	**87.51**	85.07	85.10
Llama-3.3	0-shot	79.08	79.08	**81.05**	81.05	**85.53**	**85.53**	85.67	85.67	81.69	82.69
	CoT	79.64	**81.29**	79.57	80.79	83.42	84.51	85.20	86.13	81.89	83.08
(GT)	Expert-CoT	79.22	80.32	80.68	**81.77**	83.68	85.15	85.75	86.24	**82.42**	**83.35**
(Filter)	Expert-CoT	**80.20**	81.24	80.24	81.16	82.63	83.51	**85.83**	**86.27**	82.40	83.18
OpenBioLLM	0-shot	51.60	51.73	54.18	54.26	61.58	61.71	63.66	63.74	57.67	57.76
	CoT	53.02	53.03	54.78	54.96	62.34	62.56	63.81	63.97	59.88	60.34
(GT)	Expert-CoT	**55.31**	**55.89**	**57.82**	**57.87**	64.66	65.03	**65.77**	**65.81**	**61.92**	**62.03**
(Filter)	Expert-CoT	54.91	54.91	55.12	55.12	**65.16**	**65.17**	65.52	65.54	61.32	61.93
OpenAI-o3	0-shot	**90.79**	**90.79**	**93.49**	**93.49**	**92.37**	**92.37**	**92.17**	**92.17**	**92.39**	**92.39**
Gemini-2.5	0-shot	87.59	87.59	89.93	89.93	90.26	90.26	89.51	89.51	89.36	89.36
		EMSQA—Private									
Qwen3-32B	0-shot	84.81	84.93	84.55	85.28	85.55	86.47	85.27	86.09	85.11	85.89
	4-shot	86.60	86.72	85.41	85.95	86.60	87.30	85.54	86.24	85.48	86.13
	8-shot	85.26	85.39	84.62	85.20	85.91	86.64	85.71	86.42	85.39	86.07
	16-shot	85.78	85.95	84.98	85.55	86.11	86.84	85.68	86.41	85.52	86.19
	32-shot	82.58	82.95	81.80	82.84	83.69	85.12	82.92	84.26	82.22	83.41
	64-shot	87.71	87.90	85.75	86.74	87.52	88.89	87.13	88.19	86.22	87.26
	CoT	91.81	92.01	87.86	89.08	90.11	91.66	89.83	91.24	88.78	90.13
(GT)	Expert-CoT	**94.42**	**94.74**	**88.63**	89.85	**92.01**	**93.58**	**91.59**	**92.83**	**89.73**	90.98
(Filter)	Expert-CoT	91.66	92.97	88.42	**90.05**	91.60	93.08	91.11	92.60	89.50	**91.20**
Llama-3.3	0-shot	75.95	76.12	77.93	78.63	76.87	77.66	76.75	77.47	78.06	78.77
	CoT	87.57	88.16	84.65	85.86	86.88	88.04	85.93	87.00	85.16	86.35
(GT)	Expert-CoT	89.72	90.48	85.90	**87.04**	**88.66**	**89.61**	87.55	88.58	86.49	87.62
(Filter)	Expert-CoT	**90.17**	**90.70**	**86.06**	86.98	88.41	89.47	**87.64**	**88.63**	**86.63**	**87.65**
OpenBioLLM	0-shot	60.98	61.24	60.96	61.74	66.25	67.23	67.32	68.31	63.86	64.76
	CoT	62.16	62.73	64.35	64.89	69.85	70.71	72.13	72.42	67.01	67.77
(GT)	Expert-CoT	**63.60**	**63.83**	**65.01**	**65.88**	**70.87**	**71.84**	**73.69**	**74.74**	**68.75**	**69.82**
(Filter)	Expert-CoT	62.26	62.43	64.31	64.78	70.01	70.78	72.01	72.23	67.79	68.32

Table 14: Zero-/few-shot, CoT and Expert-CoT results for LLMs on EMSQA.

Acc, but combining both attributes yields the best improvements of up to 5.29% in the 4B model. Besides, "Subject Area" helps slightly more than "Certification" in performance improvement.

Small language models benefit more from expertise guidance. Injecting expertise labels yields only marginal improvements for the 32B LLM model, but leads to significant gains for the 4B model. This disparity suggests that larger models have likely internalized much of the relevant expertise during pre-training, whereas smaller models rely more heavily on explicit expertise guidance to steer their limited capacity toward the correct reasoning trajectory.

A.9 Error Analysis

A.9.1 Error Types

We conducted an error analysis on our best 4B model (ExpertRAG-GT + Expert-CoT), which achieved 82.24% accuracy on EMSQA public split. From the 661 errors on the test set, we randomly selected 50 samples and manually examined each question, its retrieved knowledge, and the model's reasoning process. As shown in Figure 9, we attribute the errors to five main categories:

1) Reasoning Error (42.4%): Retrieved documents were relevant and were used to answer the question, but the model's reasoning was incorrect.

2) Retrieval Error(20.3%): No relevant knowledge was re-

Model	Description	Public		Private	
		Acc	F1	Acc	F1
Qwen3-32B	CoT	84.96	84.97	88.78	90.13
	+ Subject Area	85.36	85.48	88.94	90.61
	+ Certification	85.40	85.45	88.89	90.62
	Expert-CoT	85.70	85.71	89.73	90.98
Qwen3-4B	CoT	72.35	73.09	70.58	72.02
	+ Subject Area	76.32	77.04	73.82	74.53
	+ Certification	74.53	75.37	73.29	74.31
	Expert-CoT	77.26	78.38	74.32	75.77

Table 15: Ablation Study: Effect of Expertise (Certification, Subject Area) on Expert-CoT

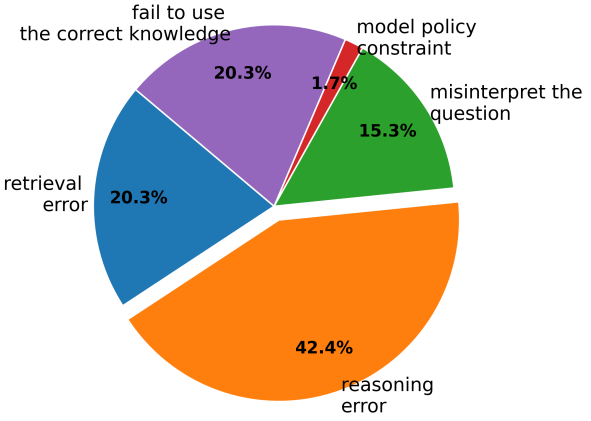

Figure 9: Error Analysis

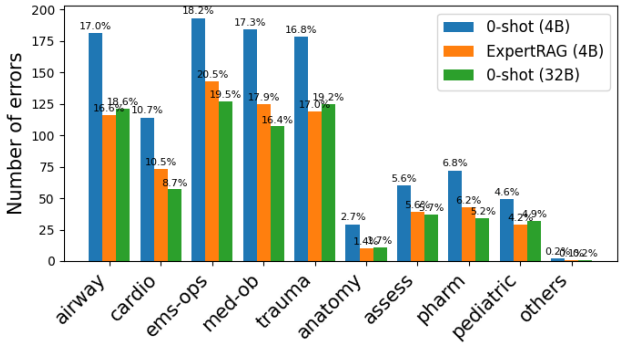

Figure 10: Error Rate per Subject Area

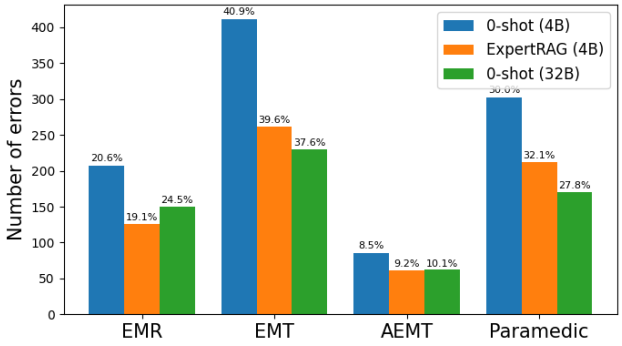

Figure 11: Error Rate per Certification

trieved, causing the model to rely on internal "model knowledge", which led to incorrect answer.

3) Knowledge Use Failure (20.3%): Part of the retrieved knowledge was relevant to answer the question but the model used other irrelevant knowledge in its reasoning.

4) Question Misinterpretation (15.3%): The model misunderstood the question. For example, for the question *what is the age cutoff for using an AED on a patient?* with choices: *a. 18 year old; b. 5 year old; c. 25 year old; d. 1 year old*, the question refers to minimum *patient* age (*Correct answer: d. 1 year old*) for routine AED use. However the model interpreted it as the minimum age of the *responder* allowed to operate an AED and incorrectly selected (*a. 18 year old*).

5) Model Policy Constraint (1.7%): The model determined that none of the answer choices were valid and abstained from responding.

A.9.2 Errors per Subject Area and Certification

To investigate how errors are distributed across subject areas and certification levels, we compare the error rates of three models, (Qwen3-4B (0-shot), ExpertRAG-4B (Expert-CoT) and Qwen3-32B (0-shot)). The results are shown in Figure 10 and Figure 11. In subject areas, all three models make most of their mistakes in the core NREMT domains: "airway", "EMS operations", "medical–OB/GYN", "trauma", and "pediatrics". Together, these categories account for the majority of errors, while "anatomy", "assessment", "pharmacology", and "others" contribute relatively few. Across certification levels, all models exhibit the highest error rates at the EMT level, followed by Paramedic and EMR, with AEMT contributing the fewest errors.

ExpertRAG-4B consistently reduces the absolute number of errors compared to the Qwen3-4B baseline across nearly all subject areas and certification levels. The Qwen3-32B has the fewest errors across three models. One finding is although the total number of errors are reduced in ExpertRAG-4B and Qwen3-32B models, the relative distribution of errors across subject areas is still the same for each model.