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, Charlottesville, VA, USA
[2]School of Medicine, University of Virginia, Charlottesville, VA, USA
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
Extended version — https://arxiv.org/pdf/2511.10900

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), 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-

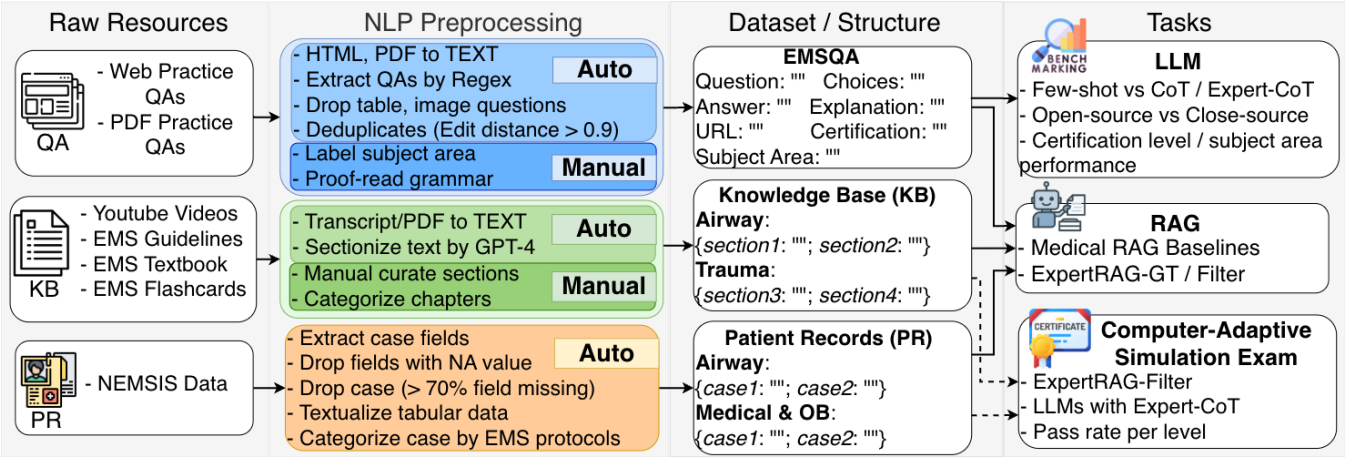

Figure 1: Overall Approach. (1) EMSQA, KB, PR Construction; (2) Tasks: Benchmark LLMs, RAG and Certification Exams.

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. 2025), 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.

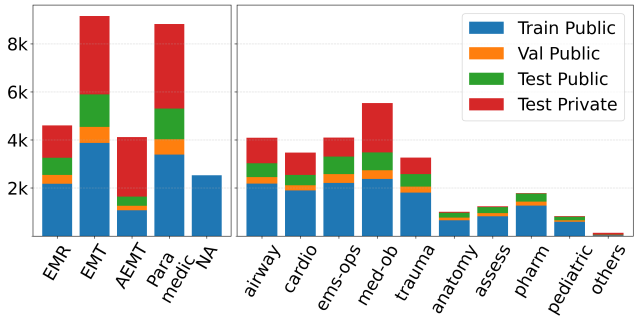

Figure 2: Distributions of Questions by Certification Level (Left) and Subject Area (Right).

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-

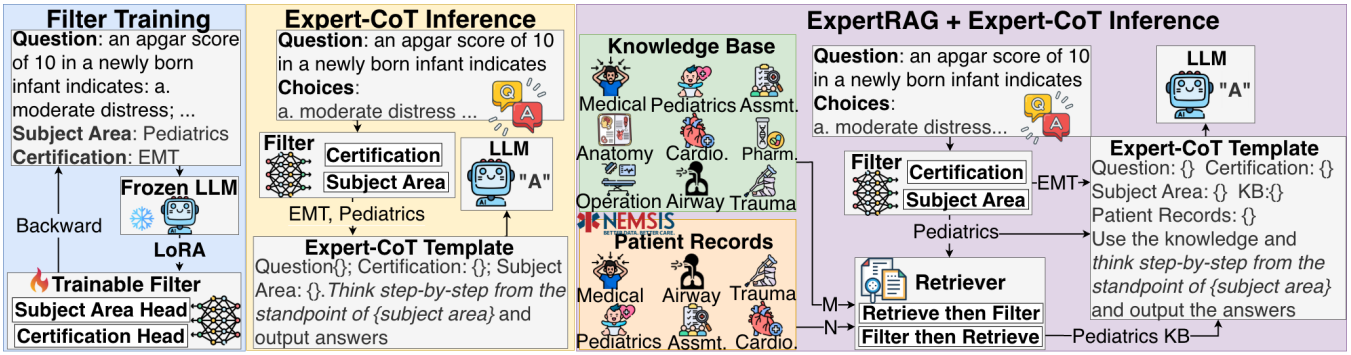

Figure 3: Expertise-Guided LLM Framework: Filter training, Expert-CoT Inference, and ExpertRAG Inference.

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:

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

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. 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:

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

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:

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

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:

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

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).

To 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,

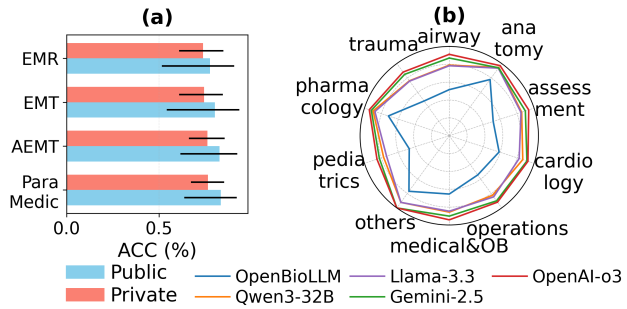

Figure 4: **(a)** Certification level 0-shot performance. **(b)** Subject area 0-shot performance on the Public dataset

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.

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.

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.

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*.

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

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

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.

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. 2025. StructRAG: Boosting Knowledge Intensive Reasoning of LLMs via Inference-time Hybrid Information Structurization. In *The Thirteenth International Conference on Learning Representations*.

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.

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

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

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.

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.

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.

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.

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.