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

Xueren Ge, Sahil Murtaza, Anthony Cortez, Homa Alemzadeh
University of Virginia
Contact: zar8jw@virginia.edu

AAAI 2026

Introduction

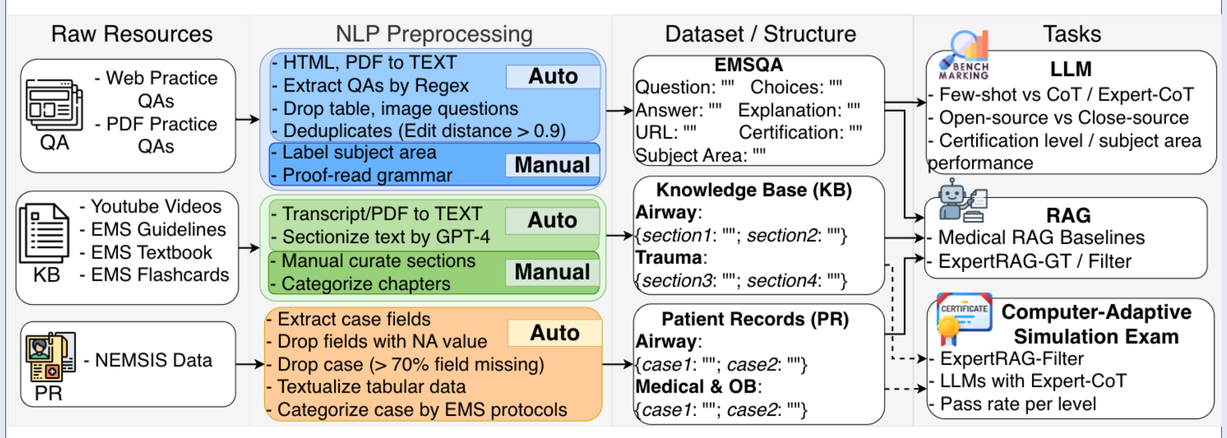
❖ Motivation

- **High-Stake Need**: Emergency Medical Service decisions require reliable, expert-level reasoning
- **Method Gap**: Current CoT/RAG treat reasoning/retrieval as generic, ignoring which expertise (subject area + certification) should guide the process.
- **Data Gap**: Existing medical QA lack structured expertise annotations and EMS-specific curated knowledge, making expert-aligned evaluation hard.

	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

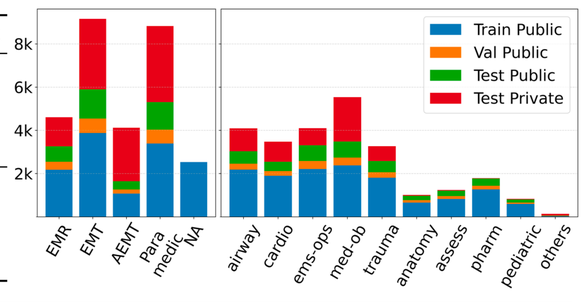
❖ Contribution

- **EMSQA**: 24.3K EMS-related QAs, covering 10 subject areas and 4 certifications. Accompanied by **EMSKB** with 40K documents and 4M Real-world patient care reports
- **Expertise-guided LLM framework**, infer EMS expertise attributes and inject them via **1) Expert-CoT** (expert-style reasoning prompts) and **2) ExpertRAG** (expert-aligned retrieval from EMS KBs + patient records).
- Expert-CoT + ExpertRAG improves accuracy by up to **4.59%**; 32B expertise-augmented models **pass all EMS certification simulations.**



EMSQA, EMSKB and Patient Records

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



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

❖ EMSQA Statistics

- 18,602 public and 5,669 private NREMT practice questions
- 4 Certification levels, 10 Subject Areas

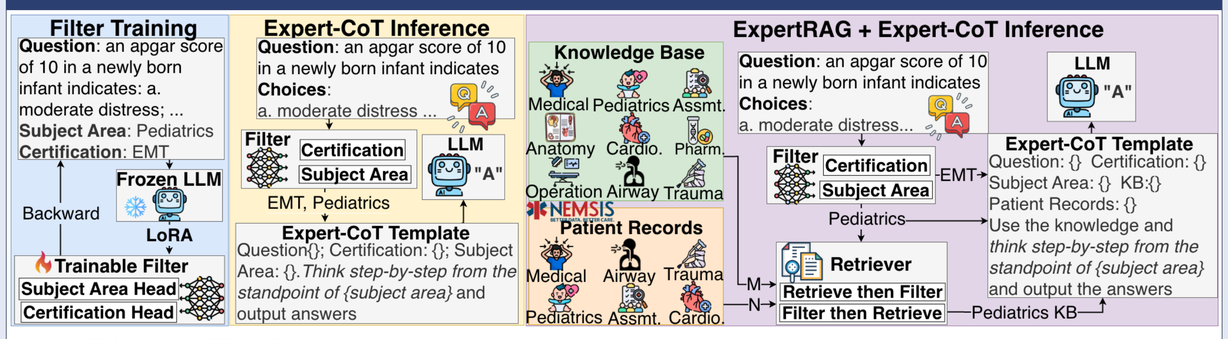
❖ EMSKB

- Crawled from 16 education resources
- Knowledge Coverage
  - Syntactic (Embedding Similarity)
  - Semantic (Vocab, EMS Concept)

❖ Patient Records

- NEMSIS Tubular Data → Textual Data
- Knowledge Coverage (Syntactic, Semantic)

Expert-CoT & ExpertRAG



❖ Filter Training

- **Input**: Q + <classify> **Output**: Subject Area; Certification

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

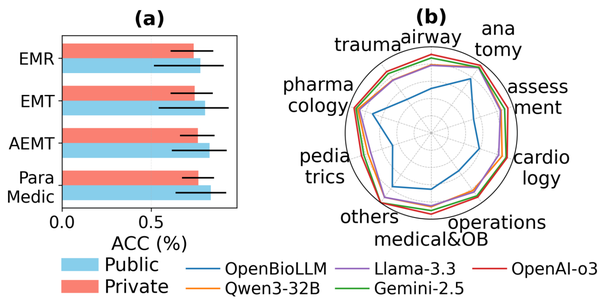
- Given the question q_i, options O_i, certification level l_i, Think from EMS-specific standpoint of subject area s_i

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

- **Global (baseline):** Retrieve top M from KB and N from PR
- **Filter then Retrieve:** Filter documents in KB & PR using subject area s_i, then retrieve top-M/N documents
- **Retrieve then Filter:** First retrieve 10×M/N candidates from KB/PR, then filter by s_i to keep top-M/N docs

Results

❖ Where do SOTA LLMs shine or stumble across subject areas and certification?



❖ What is the Expertise Classification Performance?

Split	Model	Method	Subject Area miF	Subject Area maF	Certification miF	Certification 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

❖ How much does explicit expertise injected by Expert-CoT and ExpertRAG lift baseline accuracy?

Model	Prompt	Public Acc	Public F1	Private Acc	Private 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	–	–

Model	Description	Public Acc	Public F1	Private Acc	Private 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**

❖ Can expertise-aware LLMs pass the NREMT standardized tests at different certification levels?

Model	Description	EMR Pass	EMR Score	EMR Acc	EMR T	EMT Pass	EMT Score	EMT Acc	EMT T	AEMT Pass	AEMT Score	AEMT Acc	AEMT T	Paramedic Pass	Paramedic Score	Paramedic Acc	Paramedic 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

Acknowledgements

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).