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

Xueren Ge, Sahil Murtaza, Anthony Cortez, Homa Alemzadeh

University of Virginia, Charlottesville, VA 22903 USA

{zar8jw, vpn9ej, aec3gp, ha4d}@virginia.edu

UNIVERSITY of VIRGINIA ENGINEERING

LINK LAB

UVA DSA

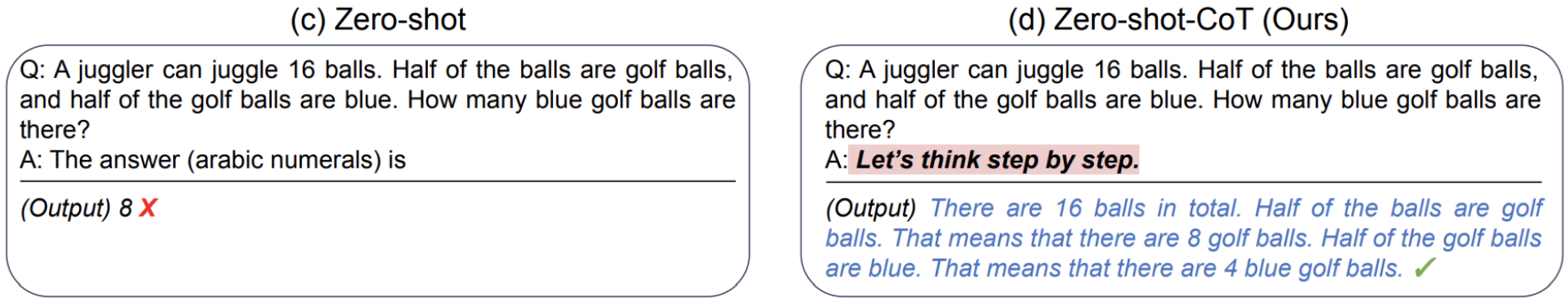
Introduction

- The advancement of Large Language Models (LLM) brought new possibilities to medical domains. particularly in the context of multiple-choice question answering (MCQA). However, important gaps remain between their current capabilities and the reasoning processes used by trained medical professionals
  - LLM sees a question and reasons or retrieves documents directly
  - Medical professionals typically begin by identifying the category of question—e.g., trauma —and then reason from that domain-specific perspective, using knowledge and protocols appropriate to their level of certification

LLM reasoning

- Chain-of-thought: step-by-step thinking
  - But think from where?

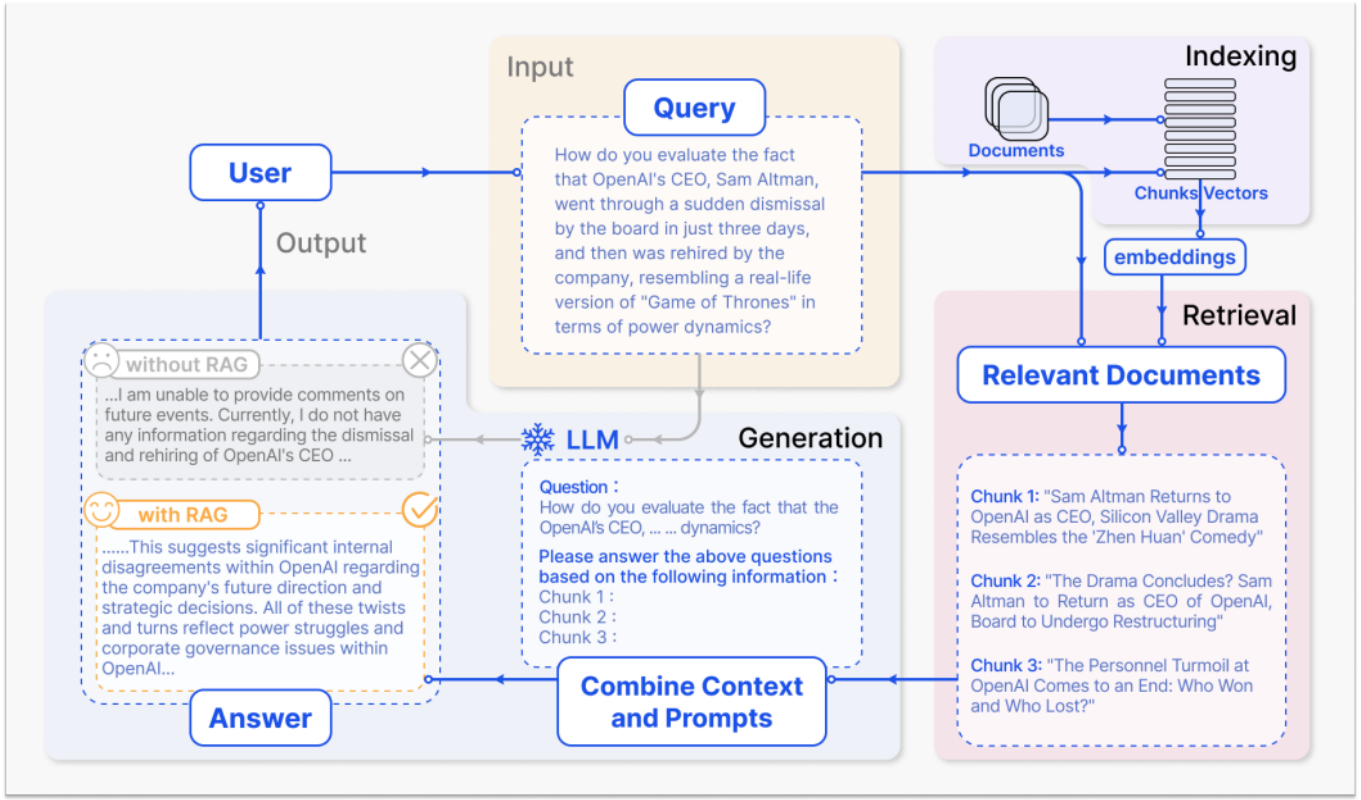
- Medical Professionals will reason from domain-specific aspects
  - EMT question → think from EMT perspective
  - Airway problem → think from airway perspective



Kojima, Takeshi, et al. "Large language models are zero-shot reasoners." *Advances in neural information processing systems* 35 (2022): 22199-22213.
Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.

Retrieval Augmentation Generation

- Standard RAG search from all documents
- Medical Professionals search from category-specific documents
  - Airway problem → Airway Documents



Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." *arXiv preprint arXiv:2312.10997* 2.1 (2023).

Medical Retrieval Augmentation Generation

- MedRAG: First pipeline to benchmark RAG in medical domain
- i-MedRAG: Introduces an iterative method to keep retrieving relevant documents
- Self-BioRAG: Introduces special tokens [REL], [USE] to train model to learn when to do RAG and judge the relevance of documents
- RAG^2 : Use rationale-based queries to retrieve documents and train a rationale-guided filter to eliminate unhelpful information
- EXPRAG: leverages historical patient records as its knowledge base
- ClinicalRAG: Proposes medical entity as the query to search document

[1] Xiong, Guangzhi, et al. "Benchmarking retrieval-augmented generation for medicine." *Findings of the Association for Computational Linguistics ACL 2024*. 2024.
[2] Xiong, Guangzhi, et al. "Improving retrieval-augmented generation in medicine with iterative follow-up questions." *Biocomputing 2025: Proceedings of the Pacific Symposium*. 2024.
[3] Jeong, Minbyul, et al. "Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models." *Bioinformatics* 40.Supplement_1 (2024): i119-i129.
[4] Rationale-Guided Retrieval Augmented Generation for Medical Question Answering (Sohn et al., NAACL 2025)
[5] Ou, Justice, et al. "Experience Retrieval-Augmentation with Electronic Health Records Enables Accurate Discharge QA." *arXiv preprint arXiv:2503.17933* (2025).
[6] Lu, Yuxing, Xukai Zhao, and Jinzhuo Wang. "ClinicalRAG: Enhancing Clinical Decision Support through Heterogeneous Knowledge Retrieval." *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*. 2024.

Medical Question Answering Datasets

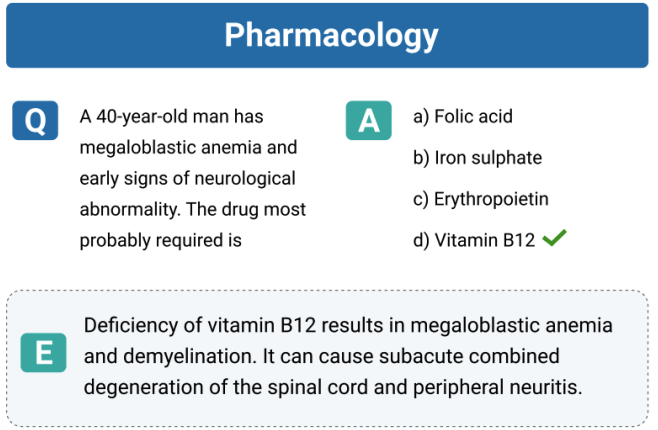
- Retrieval-based task
  - Answer is in Context
- Open-domain multiple-choice
  - require models to select the best option from choices list based on world knowledge
- Existing Medical QA benchmarks lack expertise annotations – certification levels and categories, and associated category-specific knowledge base

Record Date: 08/09/98

08/31/96 ascending aortic root replacement with homograft with omentopexy. The patient continued to be hemodynamically stable making good progress. Physical examination: BMI: 33.4 Obese, high risk. Pulse: 60. resp. rate: 18

Question: Has the patient ever had an abnormal BMI?
Answer: BMI: 33.4 Obese, high risk
Question: When did the patient last receive a homograft replacement ?
Answer: 08/31/96 ascending aortic root replacement with homograft with omentopexy.

Figure 1: Question-Answer pairs from emrQA clinical note.



	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

Contributions

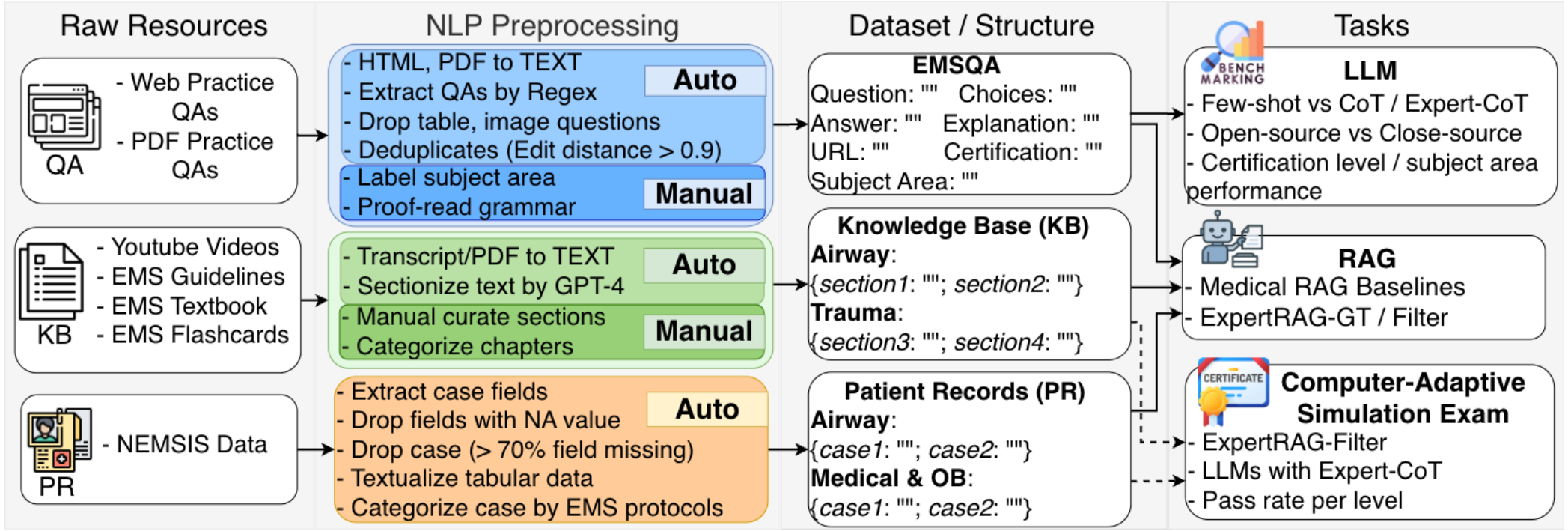
1. We introduce **EMSQA**, the first EMS MCQA dataset of 24.3K questions, curated based on public and private sources, covering 10 subject areas and 4 certification levels, and accompanied by a structured, subject area-aligned EMS knowledge base (KB) with 40K documents and 4M real-world patient care reports.

2. We propose an *expertise-guided LLM framework* that infers the domain expertise attributes and injects them into LLMs using two approaches
   - **Expert-CoT:** an expertise-guided prompting that encourages step-by-step reasoning from a domain-specific perspective
   - **ExpertRAG**—a RAG framework that selectively retrieves expertise-aligned knowledge from curated EMS knowledge bases and patient records

3. We **benchmark** multiple LLMs on EMSQA, evaluating performance across certification and subject areas, and compare against SOTAs. 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.

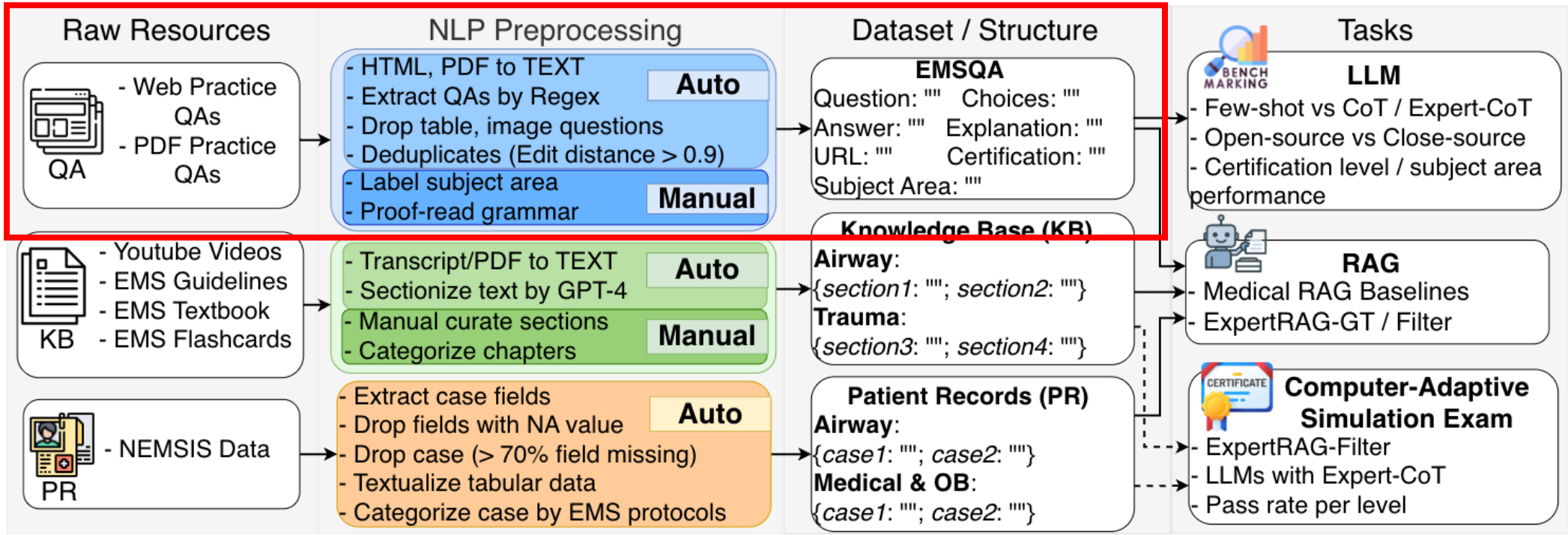
Overview

- **Dataset, Knowledge Base, Patient Records Preprocessing**
- **Downstream Tasks**
  - Benchmarks, Expert-CoT, ExpertRAG, Simulation Exam



EMSQA - Preprocessing

- Web crawling
- NLP-Preprocessing

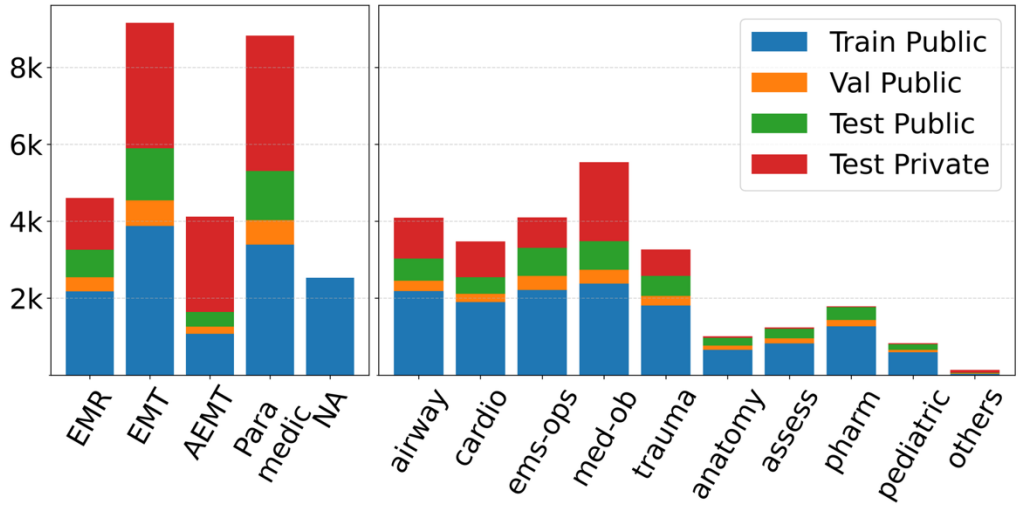


EMSQA

- Dataset statistics
  - **18,602** *public* and **5,669** *private* NREMT practice questions
  - 4 Certification levels from EMR to Paramedic
  - 10 Subject Areas

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

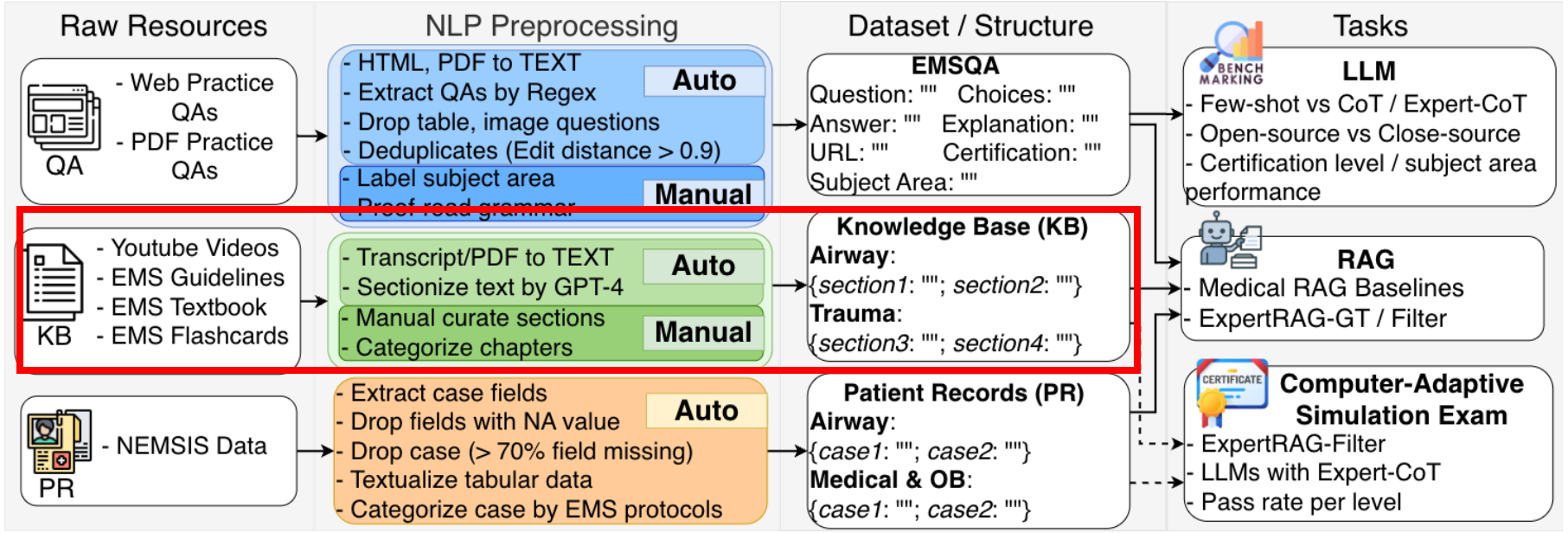


EMS Knowledge

- 16 education resources
  - Youtube Transcripts
  - Textbook/Slides
  - Medical Guidelines
  - Flashcards
- Knowledge Coverage
  - Syntactic
    - Embedding similarity
  - Semantic
    - Vocab; EMS Concept

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.

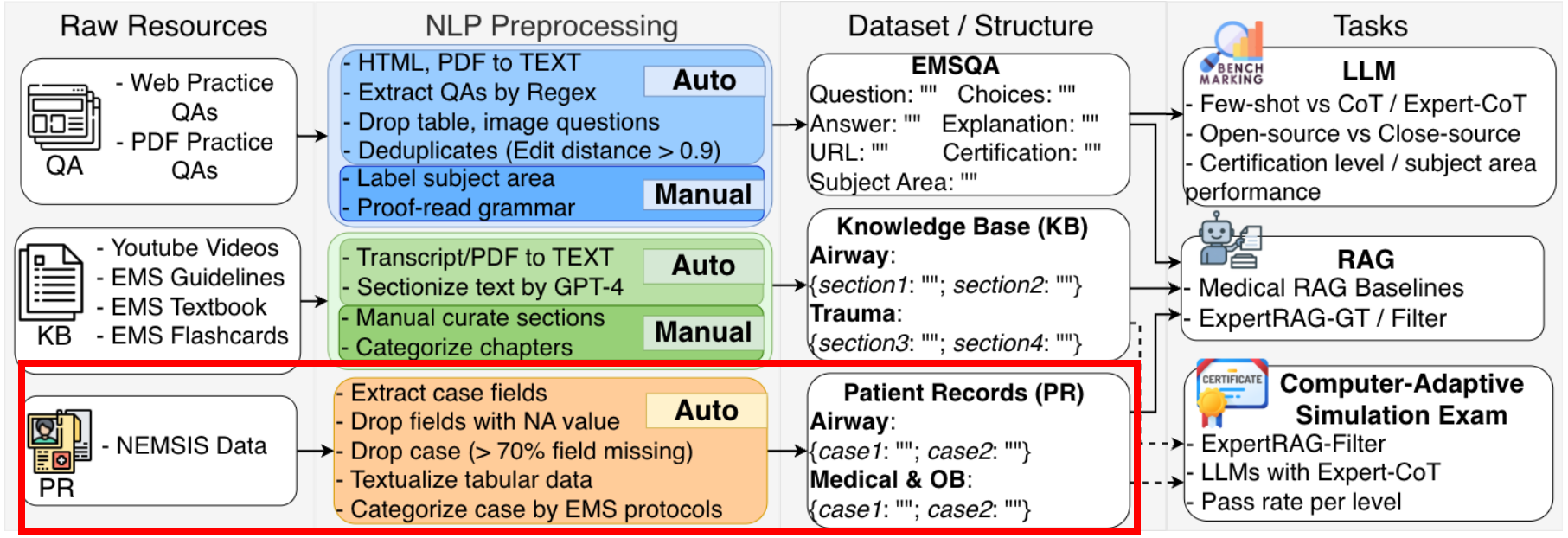


Patient Records - NEMSIS

- Tabular Data (Field, Value)
  - Drop Fields with NA Value
  - Drop case (> 70% field missing)
  - Textualize tabular data
  - Categorize Patient Records by EMS protocols

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.



Patient Records - Field

 # 1. Dispatch & Arrival
"Date - Time of Symptom Onset",

 # 2. Patient Demographics & Scene Info
"Chief Complaints",
"Possible Injury",
"Cause of Injury",
"Chief Complaint Anatomic Location",
"Chief Complaint Organ System",
"Gender",
"Age",
"Age Unit",

 # 3. Initial Assessment & Impressions
"Level of Responsiveness (AVPU)",
"Primary Symptoms",
"Other Associated Symptoms",
"Primary Impressions",
"Secondary Impressions",

 # 4. Protocols & Triage
"Protocol Age Category",
"Protocols",

 # 5. Vital Signs (in typical sequence)
"Date - Time Vital Signs Taken",
"ECG Type",
"SBP (Systolic Blood Pressure)",
"Heart Rate",
"Respiratory Rate",
"Pulse Oximetry",
"Blood Glucose Level",
"End 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","

 # 6. Medical Interventions (Medications)
"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",

 # 7. Procedure Interventions
"Date - Time Procedure Performed",
"Procedure",
"Number of Procedure Attempts",
"Response to Procedure",
"Role - Type of Person Performing the Procedure",

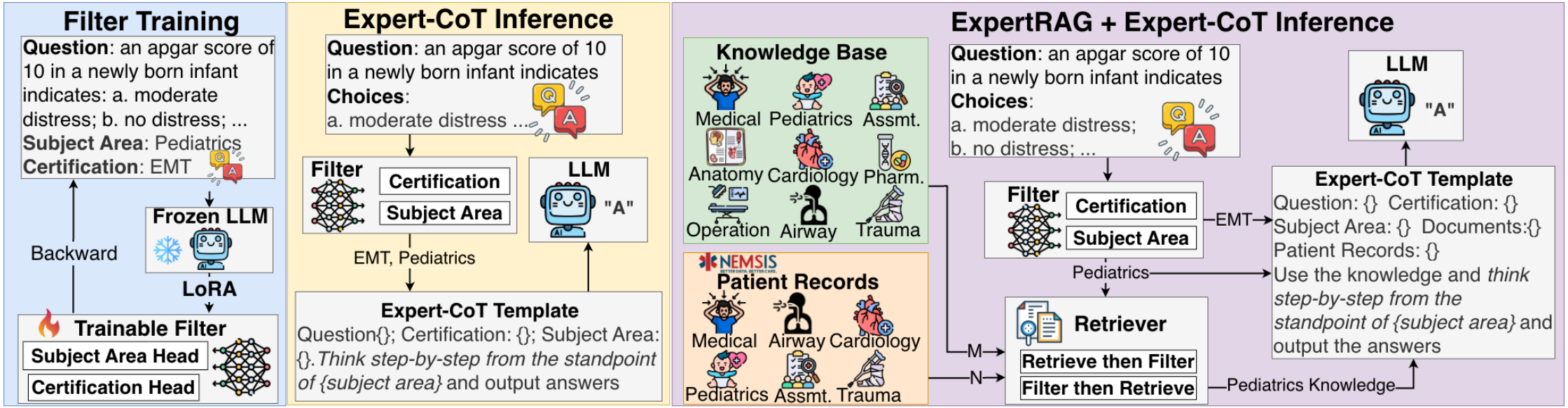
 # 8. Special Assessments / History
"Alcohol - Drug Use Indicators",
"Barriers to Patient Care",

 # 9. Cardiac Arrest Events
"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",

 # 10. Transport & Disposition
"Incident - Patient Disposition",
"EMS Transport Method",
"Transport Mode from Scene",
"Initial Patient Acuity",
"Final Patient Acuity

Methodology

- We propose an ***expertise-guided LLM framework*** with an expertise classification module (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 LLM using two strategies:
- (1) **Expert-CoT**, a prompting strategy that encodes s_i and l_i into a prompt template to guide LLM based on question-specific expertise;
- (2) **ExpertRAG**, a subject-area-specific retriever that retrieves knowledge sources conditioned on s_i



Methodology - Filter

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

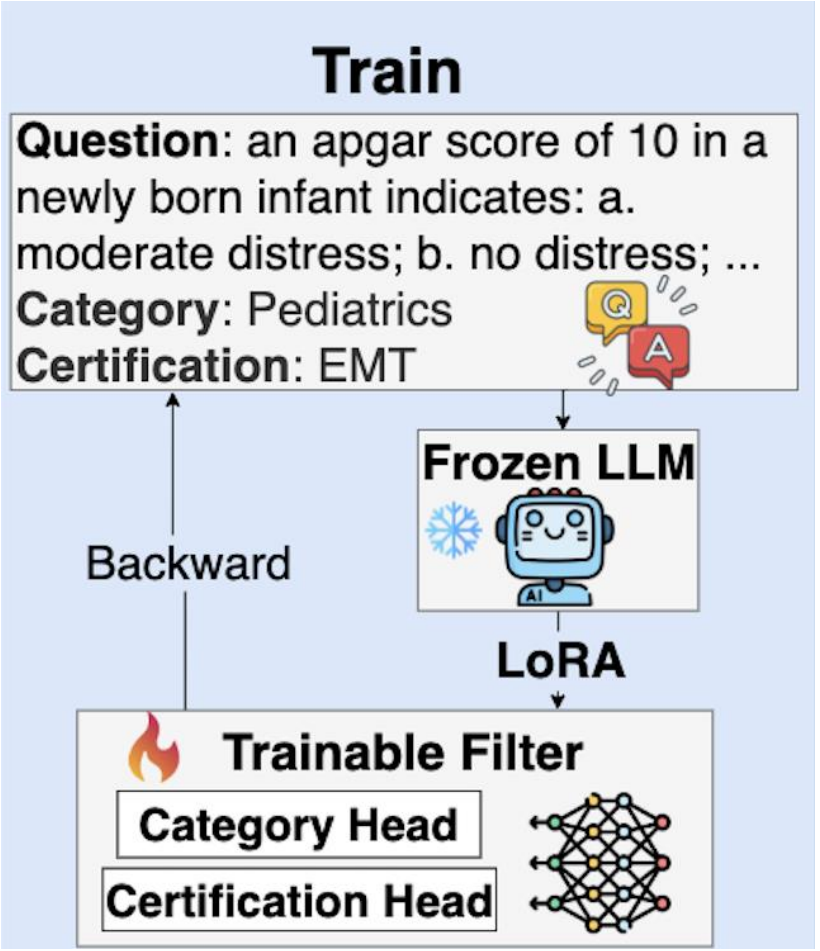
$$h_i = \mathrm{LM}_{\mathrm{last}}(q_i, \mathcal{O}_i \| \langle \mathrm{classify} \rangle),$$
$$(p_i^{\mathrm{sub}}, p_i^{\mathrm{lvl}}) = \left(\sigma(W_{\mathrm{sub}}^{\top} h_i), \sigma(W_{\mathrm{lvl}}^{\top} h_i)\right) \quad (2)$$

  - Subject Area: multi-label classification
  - Certification: multi-class classification

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

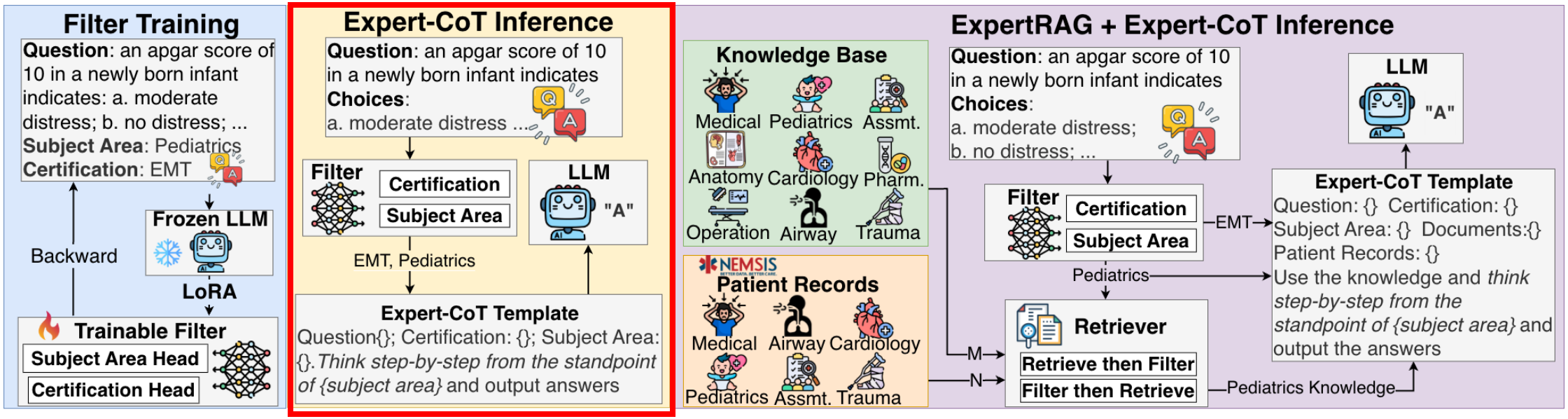
  - Inference

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



Methodology – Expert-CoT

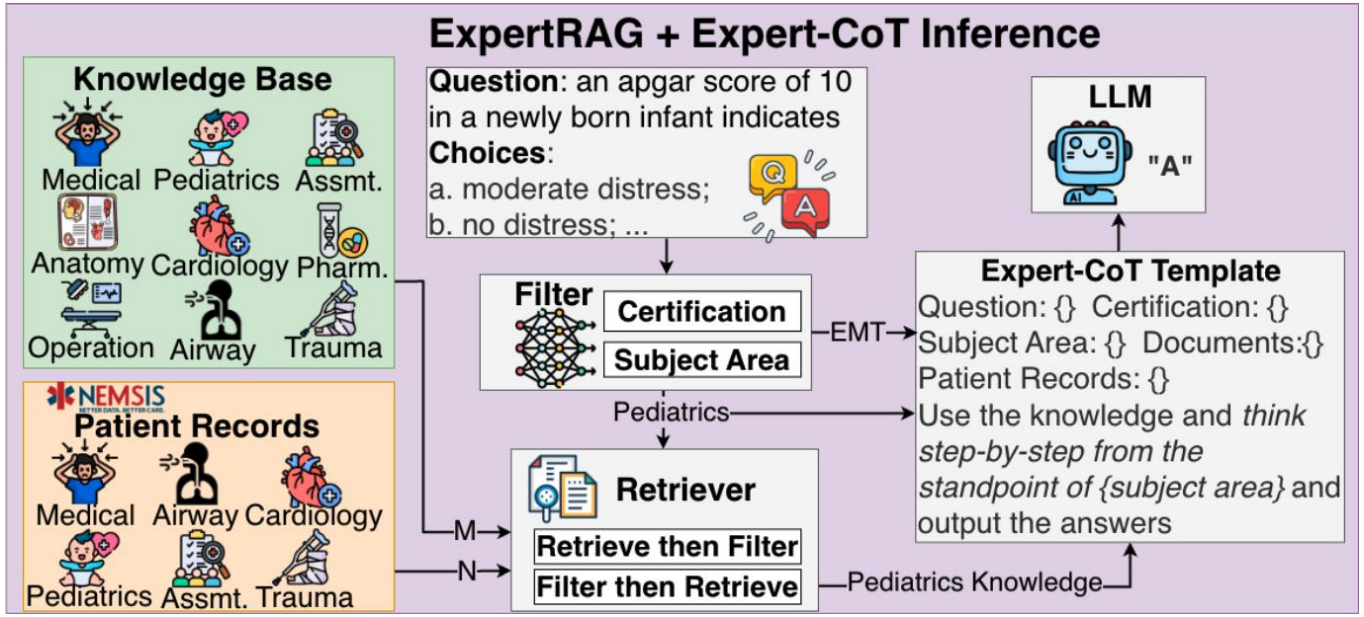
- CoT: think step by step
- Expert-CoT
  - Given the question q_i, options O_i, certification level l_i, Think from EMS-specific standpoint of subject area s_i $\quad \hat{A}_i = f^{\text{CoT-Expert}}(q_i,\ \mathcal{O}_i,\ \hat{l}_i,\ \hat{s}_i).$



Methodology - ExpertRAG

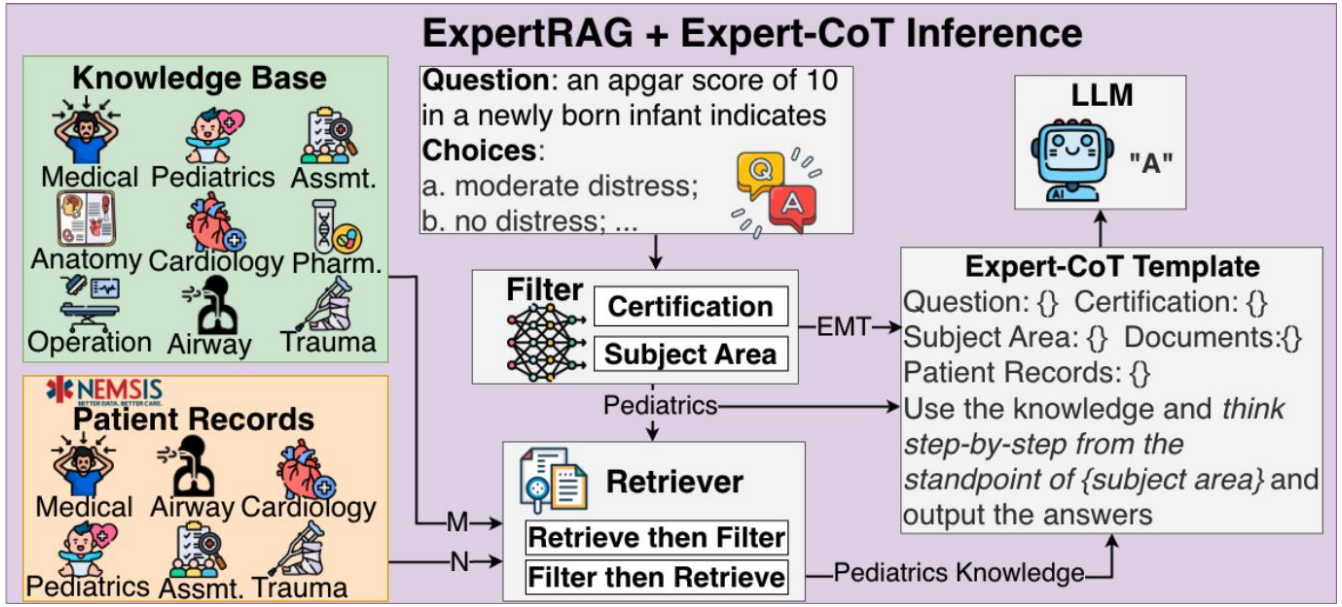
- Given a question q_i and options O_i, the **Filter** predicts each question's **subject area** $\widehat{s_i}$ and **certification level** l_i, guiding retrieval of question relevant documents $R(q_i, \widehat{s_i})$ from **knowledge base** and **patient records**. The LLM then generates the answer conditioned on the predicted **expertise** and **retrieved documents**

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



Methodology - ExpertRAG

- **Retrieval Strategy**
  - **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 from each
  - **Retrieve then Filter:** First retrieve 10×M/N candidates from KB and PR, then filter by subject area s_i to keep top-M/N relevant docs



Experiment Setup

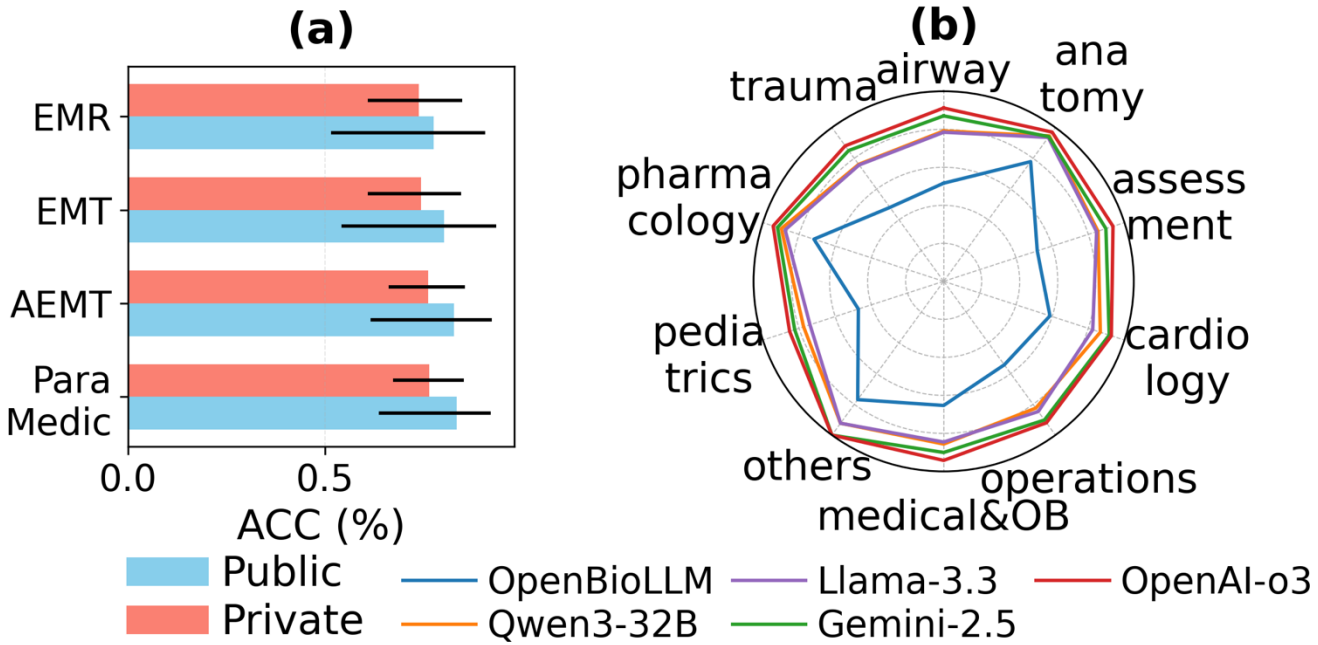
- **Baselines for LLM benchmarks**
  - Open-source LLMs (Qwen3-32B, Llama-3.3)
  - Medical LLMs (OpenBioLLM)
  - Close-source LLMs (Gemini-2.5 pro, OpenAI-o3)
- **Baselines for RAG** (Qwen3-4B)
  - Non-RAG methods (0-shot, CoT)
  - Medical RAGs (MedRAG, i-MedRAG, Self-bioRAG)
  - RAG with our collected KB and PR (RAG-KB, RAG-PR, RAG-Global)
- **Evaluation Metrics**
  - Since this is a multi-choice question answering task, we report both **exact-match Accuracy (Acc)** and **Sampled-based F1-score (F1)**

Experiment

- **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?

RQ1: Where do SOTA LLMs shine or stumble—on EMSQA across subject categories and certification levels?

- Closed-source models outperform open-source models
- Few-shot prompting improves accuracy up to a point.
- Baseline LLMs underperform on easier questions
- LLMs falter on the core NREMT domains.



Expertise Classification Performance

- The performance of our Filter vs. LLM baselines (0-shot, 4-shot, and CoT) for expertise classification
- Our Filter trained with LoRA with two classification heads significantly outperforms the baseline LLMs

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

RQ2: How much does explicit expertise injected by Expert-CoT and ExpertRAG lift baseline accuracy?

- 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.
  - 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

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.

RQ2: How much does explicit expertise injected by Expert-CoT and ExpertRAG lift baseline accuracy?

- Effect of PR and KB (2) vs. (1)
  - KB brings more improvement than PR, and combing both yields the best performance
- Effect of Expert-CoT (3) vs. (2) (and (5) vs. (4))
  - expertise-aware reasoning is better than standard reasoning
- Effect of Expert-RAG (4) vs. (2) (and (5) vs. (3))
  - Both FTR and RTF outperform global retrieval, with RTF achieving better performance.
- Effect of Filter (6) vs. (5)
  - There is a small performance drop, but ExpertRAG-Filter still outperforms the best baseline.

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		$\Delta_{Acc/F1} = +1.38/+0.46$			
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		$\Delta_{Acc/F1} = +3.35/+2.71$			
ExpertRAG	FTR	80.97	81.34	79.13	80.00
ExpertRAG	RTF	**81.11**	**81.45**	**79.17**	**80.01**
ExpertRAG-GT + Expert-CoT		$\Delta_{Acc/F1} = +4.59/+3.69$			
ExpertRAG	FTR	81.62	81.65	80.40	81.02
ExpertRAG	RTF	**82.24**	**82.26**	**80.51**	**81.16**
ExpertRAG-Filter + Expert-CoT		$\Delta_{Acc/F1} = +3.44/+2.59$			
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.

RQ3: Can expertise-aware LLMs pass the NREMT standardized tests at different certification levels?

- We subscribe MedicTest (NREMT Computer Adaptive Simulation Test)
  - 80-150 adaptively selected questions and must be completed within 2.5 hour, the NREMT cognitive exam is scored on a 100–1500 scale, with 950 as the passing threshold.
- ExpertRAG-32B with the RTF retrieval strategy achieved the highest overall score across certifications. Although the smaller LLM did not pass, it gained the most from expertise augmentation, showing the largest accuracy improvement and scoring near or above the passing threshold.

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).

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.*

Sahil Murtaza
vpn9ej@virginia.edu

Anthony Cortez
aec3gp@virginia.edu

Homa Alemzadeh
ha4d@virginia.edu

Thank You!

zar8jw@virginia.edu

National Institute of Standards and Technology
U.S. Department of Commerce

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).

https://uva-dsa.github.io/EMSQA/