NEURAL INFORMATION PROCESSING SYSTEMS

EHRNoteQA: An LLM Benchmark for Real-World Clinical Practice Using Discharge Summaries

KAIST AI Graduate School of AI
NAVER Cloud
EWHA WOMANS UNIVERSITY College of Medicine
SNUH SEOUL NATIONAL UNIVERSITY BUNDANG HOSPITAL

Presenter: Heeyoung Kwak
NAVER Cloud / Digital Healthcare LAB / Healthcare AI

Background

- Discharge Summaries
  - Written by healthcare professionals at the time of patient discharge
  - Essential clinical notes in Electronic Health Records (EHRs) summarizing a patient's entire hospital stay, from admission to discharge
  - Crucial for clinical decisions, especially during patient readmissions and handoffs
  - Challenge: Length and complexity hinder efficient retrieval of important patient information, particularly across multiple summaries

- LLMs in Healthcare
  - Large Language Models (LLMs) can analyze complex EHRs effectively
  - Potential as QA agents to support healthcare professionals
  - Necessity: a benchmark to assess LLM performance in handling discharge summaries

Motivation

- Limitations of Existing Clinical QA Benchmarks
  - Focus on general medical questions rather than patient-specific records
  - Existing QA datasets built on discharge summaries are often
    - Limited to single-note queries (e.g., dosage within a visit)
    - Have narrow topical focus (e.g., NER annotations like drugs or relations)

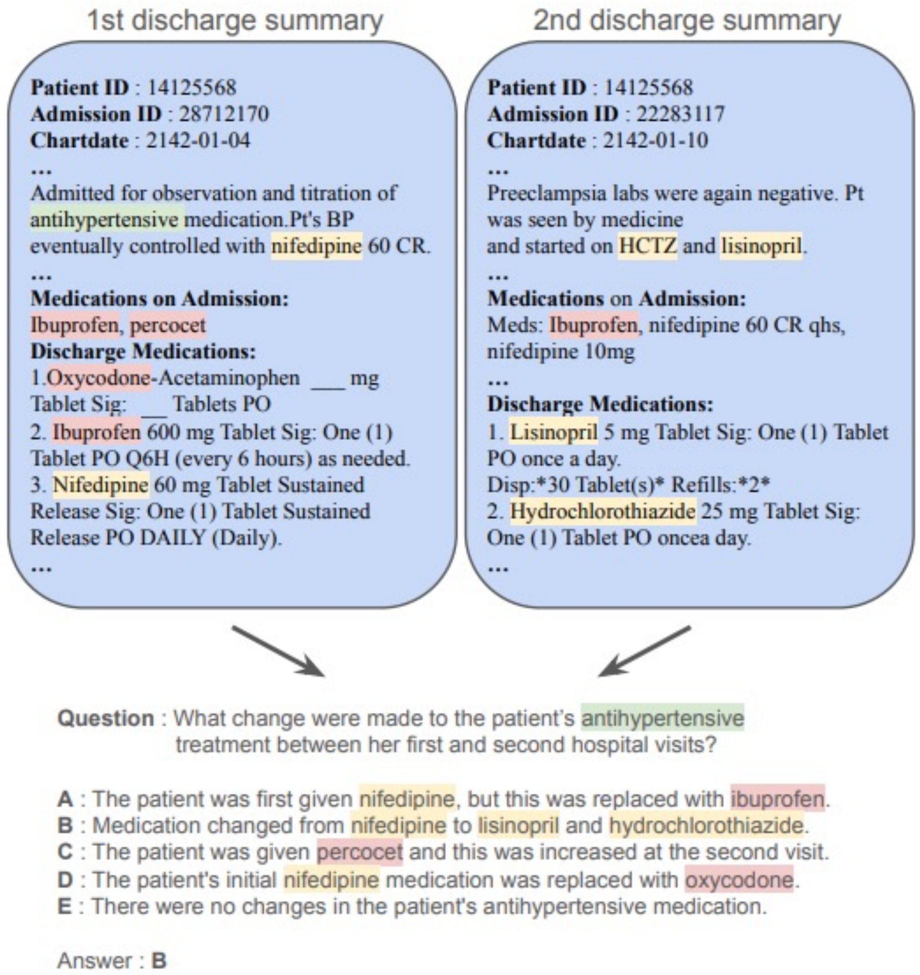
- We need an LLM benchmark that reflects real-world clinical scenarios
    - Patient-specific questions spanning multiple discharge summaries
    - Diverse clinical topics relevant to healthcare professionals

Overview - EHRNoteQA

- A novel benchmark for evaluating LLMs in real-world clinical settings

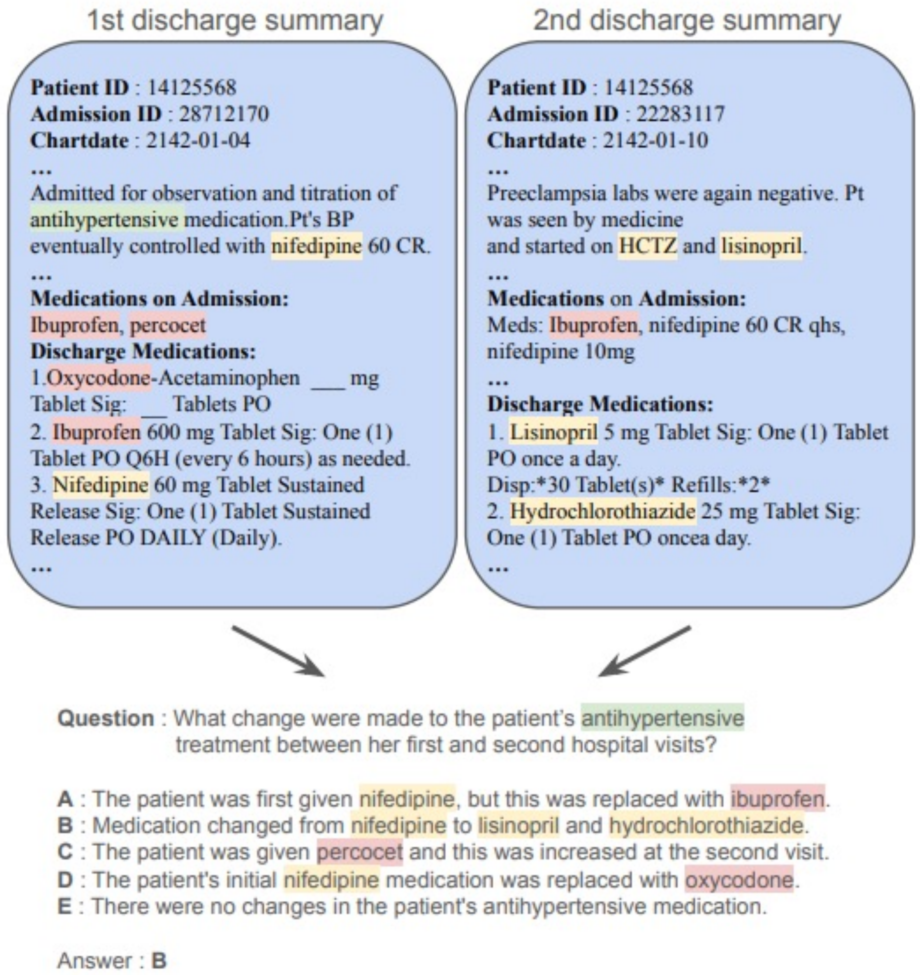
- Built on MIMIC-IV discharge summaries, featuring 962 QA pairs

- Patient-Centric:
  - Each QA pair is linked to a unique patient
  - Covers the entire sequence of their discharge summaries

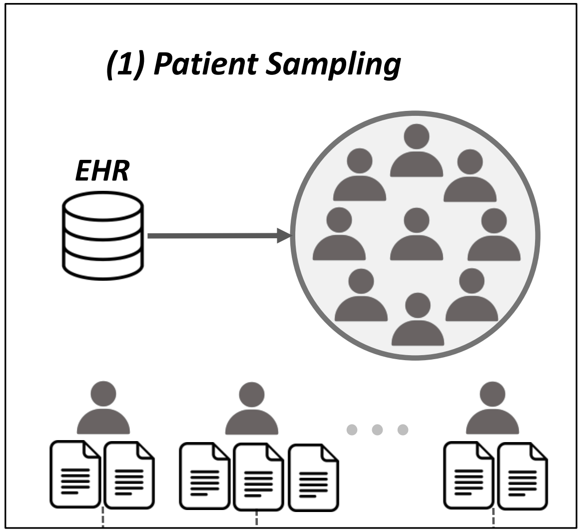


Overview - EHRNoteQA

- Real-world relevance
  - Includes multi-note queries (e.g., treatment changes)
  - Spans 10 diverse topics (e.g., treatment, vitals, history)
- Robust evaluation
  - Supports open-ended and multiple-choice formats



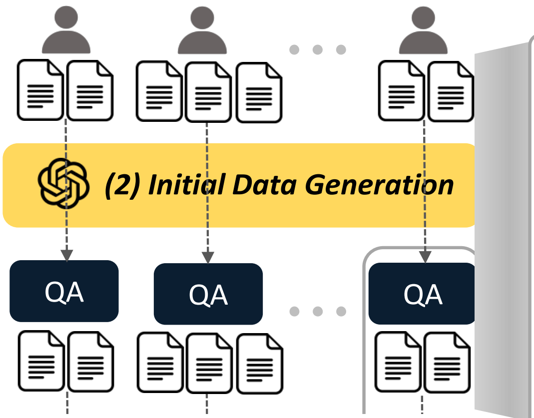
Dataset Construction – Step 1) Patient Sampling



- Patient Categorization
  - To match LLM context-length limitations:
  - **Level 1**: patients with accumulated summaries ≤ 3k tokens
    - Suitable for models handling up to 4k tokens
  - **Level 2**: patients with accumulated summaries between 3k and 7k tokens
    - Suitable for models handling up to 8k tokens
  - 1k token buffer for prompts and outputs

- Two groups cover ~70% of patients in the MIMIC-IV DB
- Sample 1,000 patients from Level 1 and 2

Dataset Construction – Step 2) Initial Data Generation



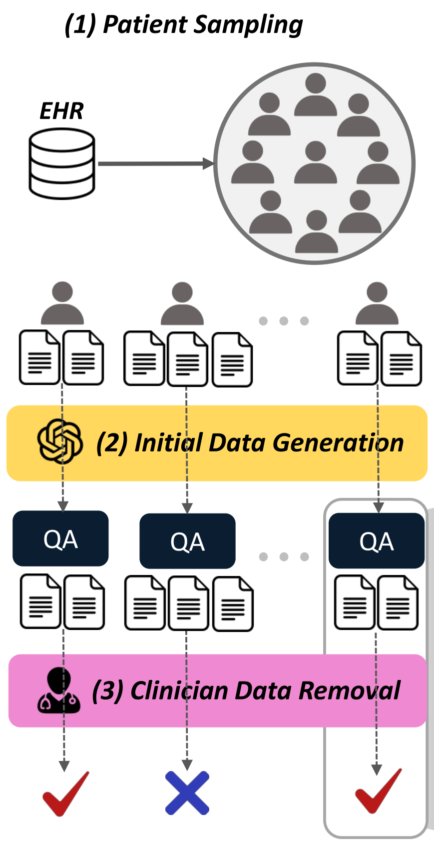
Question: What were the patient's major health issues as per the record and what was the major surgical procedure she had performed to manage these?

Correct Answer Choice: The patient's main health issues were severe aortic stenosis and atrial fibrillation, managed through an aortic valve replacement.

Incorrect Answer Choices:
- The patient was dealing with hypotension and hypertension and underwent a knee replacement surgery.
- A serious gastrointestinal bleeding was the patient's major health issue and was controlled through a midline infraumbilical laparotomy.
- The patient suffered from anxiety and depression, with a possible history of bladder cancer and was treated with a tumor excision surgery.
- Hypothyroidism and amnesia were the patient's major health concerns, and these were managed with medication.

- For each sampled patient, input the full sequence of their discharge summaries into GPT-4
- Prompt GPT-4 to generate one clinically meaningful question that a clinician may ask about the patient's discharge summary, along with its answer
- Each answer generated in two formats: multiple choice and open-ended

Dataset Construction – Step 3) Clinician Data Removal



- To ensure clinical relevance, each GPT-generated question and answer pair was reviewed by three clinicians based on the discharge summaries
- Questions that are not clinically important or unlikely to be asked by clinicians are removed

Question: What were the patient's major health issues as per the record and what was the major surgical procedure she had performed to manage these?

Correct Answer Choice: The patient's main health issues were severe aortic stenosis and atrial fibrillation, managed through an aortic valve replacement.

Incorrect Answer Choices:
- The patient was dealing with hypotension and hypertension and underwent a knee replacement surgery.
- A serious gastrointestinal bleeding was the patient's major health issue and was controlled through a midline infraumbilical laparotomy.
- The patient suffered from anxiety and depression, with a possible history of bladder cancer and was treated with a tumor excision surgery.
- Hypothyroidism and amnesia were the patient's major health concerns, and these were managed with medication.

Dataset Construction – Step 4) Clinician Data Modification

Question: What were the major health issues documented during the patient's initial admission, and what was the primary surgical procedure performed to address them?

Correct Answer Choice: The patient's main health issues were severe aortic stenosis, managed through an aortic valve replacement.

Incorrect Answer Choices:
- The patient underwent continuous positive airway pressure (CPAP) therapy for sleep apnea.
- A serious gastrointestinal bleeding was the patient's major health issue and was controlled through a midline infraumbilical laparotomy.
- The patient suffered from anxiety and depression, with a possible history of bladder cancer and was treated with a tumor excision surgery.
- The patient experienced atrial fibrillation and hypotension, both of which were addressed through aortic valve replacement.

(4) Clinician Data Modification

Question: What were the patient's major health issues as per the record and what was the major surgical procedure she had performed to manage these?

Correct Answer Choice: The patient's main health issues were severe aortic stenosis and atrial fibrillation, managed through an aortic valve replacement.

Incorrect Answer Choices:
- The patient was dealing with hypotension and hypertension and underwent a knee replacement surgery.
- A serious gastrointestinal bleeding was the patient's major health issue and was controlled through a midline infraumbilical laparotomy.
- The patient suffered from anxiety and depression, with a possible history of bladder cancer and was treated with a tumor excision surgery.
- Hypothyroidism and amnesia were the patient's major health concerns, and these were managed with medication.

- Questions that are ambiguous, overly detailed, or asks for unnecessary information were modified
- Answers were refined for accuracy and completeness
- The incorrect answer choices (multi-choice format) were modified to serve as plausible distractors

Data Statistics & Analysis

Category		MIMIC-IV		Sampled		EHRNoteQA	
Level	**# D.S.**	**# Patients**	**Avg. Length**	**# Patients**	**Avg. Length**	**Patients**	**Avg. Length**
1	1	38,926	1,819	275	1,787	264	1,812
	2	437	2,147	275	2,146	265	2,085
2	1	44,645	3,514	150	3,501	145	3,497
	2	14,176	4,470	150	4,581	144	4,520
	3	1,161	4,956	150	5,030	144	5,102
Total		99,345	-	1,000	-	962	-

Category	Example	Proportion
Treatment	What was the treatment provided for the patient's left breast cellulitis?	64%
Assessment	Was the Mitral valve repair carried out successfully?	19%
Problem	What was the main problem of the patient?	19%
Etiology	Why did the patient's creatinine level rise significantly upon admission?	20%
Sign/Symptom	What was the presenting symptom of the patient's myocardial infarction?	12%
Vitals	What was the range of the patient's blood pressure during second stay?	3%
Test Results	What were the abnormalities observed in the patient's CT scans?	14%
History	Has the patient experienced any surgical interventions prior to the acute appendicitis?	12%
Instruction	How was the patient instructed on weight-bearing after his knee replacement?	3%
Plan	What is the future course of action planned for patient's left subclavian stenosis?	5%

Experimental Settings

- Models
  - Evaluation conducted on 27 instruction-tuned LLMs, including 3 GPT models
  - Model sizes range from 7B to over 70B parameters, with various foundation models (e.g., LLaMA2, LLaMA3, Mistral, MPT and Gemma)
  - Privacy-preserving inference
    - Azure's HIPAA-compliant platform for the GPT series models
    - Local inference for the open-source LLMs
  - Use GPT-4 as an evaluator
    - GPT-4 showed high agreement with clinician evaluation

Experimental Results

Size	Model	Multi-Choice		Open-Ended		Foundation
		Level 1	Level 2	Level 1	Level 2	
	GPT4	97.16	95.15	91.30	89.61	
	GPT4-Turbo	95.27	94.23	91.30	86.61	
	GPT3.5-Turbo	88.28	84.99	82.23	75.52	
70B	Llama3-70b-Instruct	94.33	91.92	89.04	86.84	Llama3-70b
	Llama2-70b-chat	84.88	–	78.83	–	Llama2-70b
	qCammel-70	85.63	–	78.26	–	Llama2-70b
	Camel-Platypus2-70b	89.79	–	78.83	–	Llama2-70b
	Platypus2-70b-Instruct	90.36	–	80.53	–	Llama2-70b
8x7B	Mixtral-8x7b-Instruct	87.52	86.61	88.28	81.52	Mistral-7b
30B	MPT-30b-Instruct	79.96	75.52	67.11	62.59	MPT-30b-8k
13B	Llama2-13b-chat	73.65	–	70.32	–	Llama2-13b
	Vicuna-13b	82.04	–	70.51	–	Llama2-13b
	WizardLM-13b	80.91	–	74.67	–	Llama2-13b
	qCammel-13	71.46	–	66.16	–	Llama2-13b
	OpenOrca-Platypus2-13b	86.01	–	79.21	–	Llama2-13b
	Camel-Platypus2-13b	78.07	–	67.86	–	Llama2-13b
	Synthia-13b	79.21	–	74.48	–	Llama2-13b
	Asclepius-13b[1]	–	–	75.24	–	Llama2-13b
7B	Gemma-7b-it	77.50	67.21	63.71	54.27	Gemma-7b
	MPT-7b-8k-instruct	59.55	51.27	56.71	53.81	MPT-7b-8k
	Mistral-7b-Instruct	82.04	64.90	72.97	53.81	Mistral-7b
	Dolphin-2.0-mistral-7b	76.18	–	69.75	–	Mistral-7b
	Mistral-7b-OpenOrca	87.15	–	79.58	–	Mistral-7b
	SynthIA-7b	78.45	–	74.67	–	Mistral-7b
	Llama2-7b-chat	65.78	–	58.98	–	Llama2-7b
	Vicuna-7b	78.26	–	59.74	–	Llama2-7b
	Asclepius-7b[1]	–	–	66.92	–	Llama2-7b

- Performance can be affected by different factors (e.g., model size, foundation model type, instruction-tuned data, discharge summary length)
- Models struggle to perform well when handling longer/multiple discharge summaries

Reliability of EHRNoteQA as a Proxy for Clinician Evaluations

- Key question:

“Do the model scores from EHRNoteQA align with the scores given by clinicians in the targeted scenario?”

- Three clinicians evaluated LLM responses on a different set of questions
  - DiSCQ: a collection of questions asked by medical experts based on the MIMIC-III discharge summaries
  - created by medical experts not involved in EHRNoteQA

- Then, these clinician-evaluated scores were compared to the LLM scores obtained from EHRNoteQA and other benchmark datasets

Reliability of EHRNoteQA as a Proxy for Clinician Evaluations

		Clinician A		Clinician B		Clinician C	
		Spearman(ρ)	Kendall(τ)	Spearman(ρ)	Kendall(τ)	Spearman(ρ)	Kendall(τ)
Intra-Clinician correlation							
	Clinician A	-	-	0.854	0.712	0.947	0.834
	Clinician B	0.854	0.712	-	-	0.867	0.724
	Clinician C	0.947	0.834	0.867	0.724	-	-
Benchmark Comparison							
EHRNoteQA	**Open-Ended**	**0.770**	0.609	**0.805**	**0.617**	0.801	0.657
	Multi-Choice	0.766	**0.661**	0.732	0.574	**0.812**	**0.661**
Discharge Summary QA	**emrQA**	0.696	0.522	0.653	0.518	0.661	0.475
	Yue et al.	0.509	0.344	0.502	0.315	0.542	0.344
Clinical Benchmark	**MedQA**	0.590	0.453	0.497	0.354	0.683	0.535
	MedMCQA	0.672	0.512	0.505	0.378	0.737	0.594
	PubMedQA	0.122	0.100	0.071	0.059	0.167	0.088
	MMLU*	0.684	0.543	0.646	0.503	0.804	0.637
General Benchmark	**ARC**	0.534	0.425	0.522	0.373	0.583	0.460
	HellaSwag	0.284	0.206	0.247	0.177	0.373	0.265
	MMLU	0.579	0.437	0.567	0.408	0.651	0.507
	TruthfulQA	0.652	0.484	0.650	0.538	0.741	0.590
	Winogrande	0.439	0.307	0.383	0.278	0.480	0.336
	GSM8K	0.202	0.159	0.256	0.165	0.222	0.147
	AVG	0.575	0.429	0.596	0.425	0.619	0.476
Evaluation Method Comparison							
EHRNoteQA Open-Ended	**GPT-4 Eval**	**0.770**	**0.609**	**0.805**	**0.617**	**0.801**	**0.657**
	BLEU	0.155	0.112	0.037	0.059	0.014	-0.006
	ROUGE-L	0.500	0.324	0.398	0.283	0.356	0.241
	Exact Match	0.422	0.288	0.336	0.236	0.266	0.194
	SentenceBERT	0.710	0.524	0.726	0.555	0.652	0.453
	ClinicalBERT	0.536	0.382	0.552	0.389	0.394	0.288
EHRNoteQA Multi-Choice	**GPT-4 Eval**	**0.766**	**0.661**	**0.732**	**0.574**	**0.812**	**0.661**
	Probability(index)	0.622	0.472	0.596	0.444	0.676	0.519
	Probability(value)	0.514	0.437	0.523	0.456	0.549	0.437

- The experiment results show a high correlation between the clinician-evaluated LLM scores and the EHRNoteQA-evaluated LLM scores (0.6~0.8), outperforming other benchmark datasets

Reliability of EHRNoteQA as a Proxy for Clinician Evaluations

		Clinician A		Clinician B		Clinician C	
		Spearman(ρ)	**Kendall(τ)**	**Spearman(ρ)**	**Kendall(τ)**	**Spearman(ρ)**	**Kendall(τ)**
Intra-Clinician correlation							
	Clinician A	-	-	0.854	0.712	0.947	0.834
	Clinician B	0.854	0.712	-	-	0.867	0.724
	Clinician C	0.947	0.834	0.867	0.724	-	-
Benchmark Comparison							
EHRNoteQA	**Open-Ended**	**0.770**	0.609	**0.805**	**0.617**	0.801	0.657
	Multi-Choice	0.766	**0.661**	0.732	0.574	**0.812**	**0.661**
Discharge Summary QA	**emrQA**	0.696	0.522	0.653	0.518	0.661	0.475
	Yue et al.	0.509	0.344	0.502	0.315	0.542	0.344
Clinical Benchmark	**MedQA**	0.590	0.453	0.497	0.354	0.683	0.535
	MedMCQA	0.672	0.512	0.505	0.378	0.737	0.594
	PubMedQA	0.122	0.100	0.071	0.059	0.167	0.088
	MMLU*	0.684	0.543	0.646	0.503	0.804	0.637
General Benchmark	**ARC**	0.534	0.425	0.522	0.373	0.583	0.460
	HellaSwag	0.284	0.206	0.247	0.177	0.373	0.265
	MMLU	0.579	0.437	0.567	0.408	0.651	0.507
	TruthfulQA	0.652	0.484	0.650	0.538	0.741	0.590
	Winogrande	0.439	0.307	0.383	0.278	0.480	0.336
	GSM8K	0.202	0.159	0.256	0.165	0.222	0.147
	AVG	0.575	0.429	0.596	0.425	0.619	0.476
Evaluation Method Comparison							
EHRNoteQA Open-Ended	**GPT-4 Eval**	**0.770**	**0.609**	**0.805**	**0.617**	**0.801**	**0.657**
	BLEU	0.155	0.112	0.037	0.059	0.014	-0.006
	ROUGE-L	0.500	0.324	0.398	0.283	0.356	0.241
	Exact Match	0.422	0.288	0.336	0.236	0.266	0.194
	SentenceBERT	0.710	0.524	0.726	0.555	0.652	0.453
	ClinicalBERT	0.536	0.382	0.552	0.389	0.394	0.288
EHRNoteQA Multi-Choice	**GPT-4 Eval**	**0.766**	**0.661**	**0.732**	**0.574**	**0.812**	**0.661**
	Probability(index)	0.622	0.472	0.596	0.444	0.676	0.519
	Probability(value)	0.514	0.437	0.523	0.456	0.549	0.437

- Among different evaluation methods for EHRNoteQA, GPT-4 based evaluations show the highest correlation with clinician-evaluated LLM scores compared to other methods

Conclusion

- We present EHRNoteQA, a novel benchmark to evaluate LLMs in real-world clinical scenarios for answering clinicians' questions regarding discharge summaries.
- EHRNoteQA is built upon MIMIC-IV EHRs, and consists of about 1k different patient-specific QA pairs.
- EHRNoteQA questions often require information across multiple discharge summaries, and the question span a diverse set of clinical topics.
- Our experiment results validate EHRNoteQA as a reliable proxy for actual expert evaluation.