Overview of the ArchEHR-QA 2025 Shared Task on Grounded Question Answering from Electronic Health Records

Sarvesh Soni, Soumya Gayen, Dina Demner-Fushman
Division of Intramural Research
National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
sarvesh.soni@nih.gov, ddemner@mail.nih.gov

Abstract

This paper presents an overview of the ArchEHR-QA 2025 shared task, which was organized with the 24th BioNLP Workshop at ACL 2025. The goal of this shared task is to develop automated responses to patients' questions by generating answers that are grounded in key clinical evidence from patients' electronic health records (EHRs). A total of 29 teams participated in the task, collectively submitting 75 systems, with 24 teams providing their system descriptions. The submitted systems encompassed diverse architectures (including approaches that select the most relevant evidence prior to answer generation), leveraging both proprietary and open-weight large language models, as well as employing various tuning strategies such as fine-tuning and few-shot learning. In this paper, we describe the task setup, the dataset used, the evaluation criteria, and the baseline systems. Furthermore, we summarize the methodologies adopted by participating teams and present a comprehensive evaluation and analysis of the submitted systems.

1 Introduction

The volume of messages received through patient portals is on the rise, which includes requests from patients for medical information (Holmgren et al., 2023; Martinez et al., 2024). This is one of the main contributors to increasing clinician burden. One promising strategy to address this challenge is to assist clinicians in formulating responses to patient inquiries. To this end, automatically generating answers to questions from patients considering their medical records is important.

While there is extensive work on answering general health-related queries from patients (Welivita and Pu, 2023), relatively little focuses on addressing patient questions specifically about their own medical records. Within the work on patient portal messages, most research has focused on message triage (Ren et al., 2023; Liu et al., 2024c) or on helping patients formulate their questions (Liu et al., 2024b). Efforts to automatically generate answers to patient questions rarely incorporate relevant information from the patient's medical record (Liu et al., 2024a; Chen et al., 2024). Among the few that do, none evaluate how effectively the generated responses leverage that evidence (Small et al., 2024; Garcia et al., 2024).

Grounding an answer in evidence is the process of citing or referencing specific segments of the input evidence to support the generated response (Chandu et al., 2021). This practice is especially critical in medicine, where accuracy and traceability are paramount–particularly when the target users are not proficient in medical knowledge (Haug and Drazen, 2023). Although grounding has been extensively studied in open-domain (Wang et al., 2025; Sung et al., 2025), its application in the clinical domain remains relatively underexplored.

To foster research in these underexplored areas of clinical natural language processing (NLP), we introduced the ArchEHR-QA (pronounced *"Archer"*) shared task[1]. The goal of the task is to develop automated systems that generate answers to patients' questions, grounded in key clinical evidence from their electronic health records (EHRs). Participants were provided with patient-posed questions, their clinician-interpreted versions, and corresponding clinical notes. Systems were expected to produce answers accompanied by sentence-level citations to the relevant sentences of the clinical note.

2 ArchEHR-QA 2025 Task Description

Given a patient-posed natural language question, the corresponding clinician-interpreted question, and the patient's clinical note excerpt, the task is to generate a natural language answer with sentence-

[1] archehr-qa.github.io

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 396–405

Patient Question	I had severe abdomen pain and was hospitalised for 15 days in ICU, diagnosed with CBD sludge thereafter on udiliv. Doctor advised for ERCP. My question is if the sludge was there does not the medication help in flushing it out? Whether ERCP was the only cure?
Clinician Question	Why was ERCP recommended to him over continuing a medication-based treatment?
Clinical Note Excerpt	**1:** During the ERCP a pancreatic stent was required to facilitate access to the biliary system (removed at the end of the procedure), and a common bile duct stent was placed to allow drainage of the biliary obstruction caused by stones and sludge. **2:** However, due to the patient's elevated INR, no sphincterotomy or stone removal was performed. **3:** Frank pus was noted to be draining from the common bile duct, and post-ERCP it was recommended that the patient remain on IV Zosyn for at least a week. **4:** The Vancomycin was discontinued. **5:** On hospital day 4 (post-procedure day 3) the patient returned to ERCP for re-evaluation of her biliary stent as her LFTs and bilirubin continued an upward trend. **6:** On ERCP the previous biliary stent was noted to be acutely obstructed by biliary sludge and stones. **7:** As the patient's INR was normalized to 1.2, a sphincterotomy was safely performed, with removal of several biliary stones in addition to the common bile duct stent. **8:** At the conclusion of the procedure, retrograde cholangiogram was negative for filling defects.
Clinician Answer	An endoscopic retrograde cholangiopancreatography, ERCP, was recommended to place a common bile duct stent **[1]**. This stent was placed to allow drainage of the biliary obstruction which was caused by stones and sludge **[1]**. Due to no improvement in liver function, the patient needed a repeat ERCP **[5]**. The repeat ERCP showed that the biliary stent placed in the first ERCP was obstructed by stones and sludge **[6]**. The stones and stent were successfully removed during this procedure by performing a sphincterotomy **[7]**.
Sample System Answer	ERCP was recommended as medication alone may not effectively clear the biliary obstruction caused by stones and sludge **[1]**. The procedure allowed for direct visualization, stenting, and subsequent removal of biliary stones and sludge **[1,7,8]**. Medication, such as IV Zosyn, was used adjunctively to manage infection but was not sufficient to flush out the sludge and stones on its own **[3]**.

Table 1: Example annotated patient case from the ArchEHR-QA dataset. Focus areas in the patient question are underlined, which are taken into account when creating the clinician question. In the clinical note excerpt, the sentences with identifiers `[1,5,6,7]` are `essential`, `[2,8]` are `supplementary`, and `[3,4]` are `not-relevant`.

level citations to the specific clinical note sentences.

3 Data Description

We used a subset of 120 patient cases from the ArchEHR-QA dataset (Soni and Demner-Fushman, 2025). Each patient case in the dataset comprises a hand-curated, realistic patient question (reflective of patient portal messages), relevant focus areas identified within the question (as determined by a clinician), corresponding clinician-rewritten version (crafted to aid in formulating responses), and note excerpt providing essential clinical context (Table 1). The dataset was curated by aligning real patient questions posted to public health forums with clinical notes from publicly accessible EHR databases, namely, MIMIC-III and MIMIC-IV (Johnson et al., 2016, 2023). Each sentence in the note excerpt is manually annotated to mark its importance in answering the question as "`essential`" (must be cited in the answer), "`supplementary`" (may be cited to provide support), or "`not-relevant`" (should not be cited). For more details about the dataset curation process, please refer to the dataset paper (Soni and Demner-Fushman, 2025).

A total of 20 patient cases were provided to the participants with sentence relevance keys for the development and validation of systems. The remaining 100 patient cases were used for testing the participant systems and released to the participants closer to the final submission date without the sentence relevance labels. Tables 2 and 3 provide the dataset statistics.

4 Evaluation

4.1 Metrics

Submissions were evaluated based on their use of clinical evidence for grounding (*"Factuality"*) and the relevance of the generated answers (*"Relevance"*). The scoring script is available on GitHub[2].

Factuality is assessed by calculating Precision, Recall, and F1 Scores between the cited evidence sentences in the generated answers (i.e., predicted as "`essential`") and the manually anno-

[2] github.com/soni-sarvesh/archehr-qa

	Patient Question		Clinician Question		Note Excerpt		Clinician Answer	
	Dev	**Test**	**Dev**	**Test**	**Dev**	**Test**	**Dev**	**Test**
Mean	85.2	92.3	10.8	10.6	320.8	380.4	73.6	72.4
Median	81.0	74.5	10.0	10.0	320.5	345.0	74.0	73.0
S.D.	35.1	62.4	2.8	3.9	174.6	213.3	2.3	3.6
Min	40.0	33.0	7.0	3.0	109.0	76.0	66.0	55.0
Max	170.0	440.0	17.0	21.0	678.0	1028.0	78.0	76.0

Table 2: Word count statistics by dataset split. *Dev*: development; *S.D.*: standard deviation.

	Sentences	
Relevance	**Dev**	**Test**
`all`	21.4 (100%)	26.0 (100%)
`essential`	6.0 (28.3%)	6.6 (25.3%)
`supplementary`	1.3 (6.1%)	5.5 (21.3%)
`not-relevant`	14.1 (65.7%)	13.9 (53.4%)

Table 3: Average sentence counts by relevance and dataset split. *Dev*: development.

tated ground truth sentence relevance labels. Two variations of Citation F1 Scores are calculated. In the *"strict"* variation, only `essential` sentences are considered as answers. In the *"lenient"* variation, both `essential` and `supplementary` sentences are considered as answers.

Relevance is evaluated by comparing the generated answer text with the ground truth answer. Two variations of ground truth answers were used for relevance computations: clinician-authored answer and a concatenation of `essential` note sentences with patient and clinician questions. A suite of text and semantics based relevance metrics are used to compare the predicted and ground truth text: BLEU (Papineni et al., 2002), ROUGE-LSum (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2019), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023).

4.2 Baseline

As a simple yet strong baseline, we prompted the LLaMa 3.3 70B model (Grattafiori et al., 2024) in a zero-shot setting to generate an answer using both the patient and clinician questions, along with the note excerpt as input. We provided the note sentence identifiers to the model and instructed it to cite the specific sentence IDs within its generated answer. In instances where the model failed to follow the required formatting or citation guidelines, we iteratively prompted the model with specific feedback from the previous attempt (e.g., an *invalid citation*) up to five times.

5 Participation

5.1 Participating Teams

We used the Codabench platform[3] to facilitate shared task submission process (Xu et al., 2022). In total, 29 teams participated in the task and submitted a total of 75 systems. Of these, 24 teams provided a description of their submitted system. We report the evaluation scores exclusively for those submissions accompanied by a system description.

5.2 Results

Participants were provided with a preliminary version of sentence relevance keys during the development phase, where note excerpts and questions were used to compute the relevance scores. Table 4 presents the submission results on the test set (with hidden keys) using this setup. In this setting, DMIS Lab (Hwang et al., 2025) achieved the highest overall score of 53.7% with a strict micro F1 score of 58.6% and an average relevance score of 48.8%. This was followed by Neural (Bogireddy et al., 2025) and LAILab (Le et al., 2025), which attained overall scores of 51.5% and 51.0%, respectively. Notably, ArgHiTZ (Cortes et al., 2025) obtained the highest strict micro F1 score of 60.5%.

Upon completion of the annotation reconciliation process, we recalculated the evaluation metrics using revised sentence relevance keys and clinician-authored reference answers. These results are reported in Table 5. While the overall score range remained relatively stable, there were substantial changes in the ranking of individual systems, and we observed a general drop in the overall scores for

[3] codabench.org/competitions/5302

Rank			Team	Factuality						Relevance							
				Lenient			Strict			Text			Semantics				
O	F	R	ID	P	R	F1	P	R	F1	BL	RG	SA	BS	AS	MD	Avg	OS
1	4	1	DMIS Lab	61.2	59.2	60.2	57.9	59.3	58.6	14.3	46.5	36.7	53.9	92.4	49.3	48.8	53.7
2	3	2	Neural	58.4	63.7	60.9	55.4	63.8	59.3	8.5	34.1	73.1	39.1	67.3	40.0	43.7	51.5
3	2	4	LAILab	59.7	66.0	62.7	56.0	65.5	60.4	6.5	32.7	69.2	37.4	65.3	38.4	41.6	51.0
4	6	5	LAMAR	64.0	53.5	58.3	60.6	53.6	56.9	6.0	32.1	65.8	36.4	64.3	43.6	41.4	49.1
5	14	3	ssagarwal	71.7	35.6	47.6	68.8	36.2	47.5	4.7	31.1	70.0	36.9	74.9	38.0	42.6	45.0
6	7	8	LIMICS	63.6	49.6	55.8	59.9	49.4	54.2	3.0	26.2	61.2	31.2	52.3	39.4	35.5	44.9
7	10	7	cuni-a	60.2	48.1	53.5	56.9	48.1	52.1	5.1	26.5	63.2	32.0	58.2	37.7	37.1	44.6
8	1	22	ArgHiTZ	58.9	65.8	62.1	55.8	65.9	60.5	0.9	21.1	48.1	22.1	42.3	30.9	27.6	44.0
9	5	16	Loyola	51.1	70.5	59.3	48.3	70.5	57.3	2.9	25.5	54.4	26.1	42.4	30.8	30.4	43.9
10	8	11	unibuc-sd	66.5	47.2	55.2	62.7	47.0	53.8	1.4	22.2	53.3	27.5	53.4	38.2	32.7	43.2
11	15	6	SzegedAI	69.7	37.0	48.4	65.6	36.9	47.2	3.2	27.8	63.6	32.9	64.2	37.8	38.2	42.7
12	11	12	KRLabs	50.7	56.6	53.5	48.1	56.8	52.1	2.0	21.4	57.9	26.3	49.0	35.2	31.9	42.0
13	12	10	FK	70.0	37.9	49.2	66.7	38.2	48.6	2.0	25.4	54.4	28.2	55.8	36.8	33.8	41.2
14	9	20	UTSA-NLP	47.0	68.4	55.7	43.7	67.2	53.0	0.7	17.8	56.6	22.7	40.4	29.4	27.9	40.4
15	17	13	UIC	70.4	35.2	46.9	67.3	35.6	46.5	0.7	19.4	55.6	24.6	57.7	31.4	31.6	39.0
16	13	21	utsamuel	55.1	45.3	49.7	51.4	44.7	47.8	0.6	20.0	56.7	24.2	35.4	29.6	27.8	37.8
17	16	23	aehrc	55.5	42.0	47.8	52.9	42.4	47.1	0.6	19.0	48.4	22.5	41.9	30.3	27.1	37.1
18	18	19	unibuc-sb	61.7	35.9	45.4	58.7	36.1	44.7	0.6	19.9	49.0	23.9	43.0	32.4	28.1	36.4
19	20	17	HurLab	52.9	34.8	41.9	49.3	34.3	40.4	1.8	24.0	47.3	24.9	48.1	34.5	30.1	35.2
20	19	18	JUNLP	57.5	32.4	41.4	54.2	32.3	40.5	1.5	22.8	49.3	24.4	49.1	30.9	29.6	35.1
21	21	9	WisPerMed	59.1	27.1	37.1	55.4	26.9	36.2	2.0	22.6	61.0	29.5	62.3	25.9	33.9	35.0
22	22	15	DKIT	59.9	23.1	33.4	56.5	23.1	32.7	1.7	23.6	49.8	26.2	47.8	33.9	30.5	31.6
23	23	14	heiDS	71.2	16.0	26.2	67.7	16.1	26.0	0.7	18.1	53.6	22.2	61.0	29.9	30.9	28.5
24	24	24	razreshili	39.7	8.4	13.9	36.8	8.2	13.5	0.4	16.8	45.8	19.9	43.9	24.5	25.2	19.3
-	-	-	baseline	77.0	22.3	34.6	71.6	21.9	33.6	0.1	15.2	47.8	20.5	57.7	25.6	27.8	30.7

Table 4: Submission scores using the preliminary version of answer keys, with note excerpts and questions used for evaluating relevance. Factuality scores are reported at the micro level. O, F, R: Rank using Overall, Factuality (Strict F1), and the average Relevance score. ID: Team identifier; P: Precision; R: Recall; F1: F1 Score; BL: BLEU; RG: ROUGE; SA: SARI; BS: BERTScore; AS: AlignScore; MD: MEDCON; Avg: Overall Relevance Score; OS: Overall Score. All scores are percentages.

most submissions. Under this revised evaluation, LAMAR (Yoadsanit et al., 2025) achieved the highest overall score and strict micro F1 score of 46.9% and 58.8%, respectively. FK and unibuc-sd (Ghinea and Rîncu, 2025) followed closely, securing the second and third positions with overall scores of 46.6% and 45.6%, respectively.

5.3 Approaches

Table 6 summarizes the key characteristics of the systems submitted to the shared task. The majority of teams (20 [83.3%]) adopted a two-stage pipeline in which relevant evidence was first identified from the note excerpts, followed by answer generation in a subsequent stage. Several teams also incorporated additional post-generation steps, such as citation assignment (5 [20.8%]) or answer reformulation (8 [33.3%]) with an aim to further enhance the quality of responses.

All participating teams utilized language models as part of their systems. Over half of the teams (14 [58.3%]) employed proprietary models, such as OpenAI's GPT, while 11 teams (45.8%) used open-weight large language models (LLMs), e.g., Meta's LLaMA. Additionally, 9 teams (37.5%) integrated small language models (SLMs), such as BERT, into their systems. Model tuning strategies varied, with fine-tuning being the most common (6 [25.0%]), followed by few-shot learning (5 [20.8%]), the use of synthetic data (3 [12.5%]), and hyperparameter tuning (2 [8.3%]). Postprocessing steps to refine the generated answers were also reported, with some teams leveraging the language model itself for editing (5 [20.8%]) and some applying heuristic-based approaches (4 [16.7%]).

Among the top-scoring systems, answer reformulation emerged as a common component, so did the use of proprietary LLMs. Notably, the leading

Rank			Team	Factuality						Relevance							
				Lenient			Strict			Text			Semantics				
O	F	R	ID	P	R	F1	P	R	F1	BL	RG	SA	BS	AS	MD	Avg	OS
1	1	2	LAMAR	72.9	57.4	64.2	49.6	72.0	58.8	8.4	24.8	55.9	40.8	36.9	43.2	35.0	46.9
2	3	1	FK	78.8	40.3	53.3	59.1	55.6	57.3	7.4	24.6	53.9	40.2	47.9	41.6	35.9	46.6
3	2	6	unibuc-sd	75.4	50.4	60.4	53.0	65.2	58.4	4.2	21.3	53.0	41.0	34.6	41.9	32.7	45.6
4	5	5	ssagarwal	79.0	37.0	50.4	58.3	50.3	54.0	8.1	24.1	54.3	36.8	33.7	39.3	32.7	43.4
5	10	3	UIC	77.6	36.6	49.7	55.5	48.2	51.6	4.5	22.8	54.5	41.7	40.5	38.9	33.8	42.7
6	9	4	SzegedAI	78.4	39.3	52.3	54.4	50.2	52.2	7.1	23.4	54.1	39.3	35.0	39.1	33.0	42.6
7	6	10	LIMICS	71.2	52.3	60.3	46.9	63.5	54.0	5.5	22.2	54.3	38.6	25.7	39.3	30.9	42.5
8	4	18	Neural	67.5	69.3	68.4	42.8	81.0	56.0	6.3	20.7	53.1	30.6	25.9	33.3	28.3	42.2
9	7	13	LAILab	65.7	68.4	67.0	40.6	77.8	53.3	7.4	22.0	53.3	33.9	26.8	34.4	29.6	41.5
10	13	7	JUNLP	64.9	34.5	45.0	49.5	48.5	49.0	5.2	21.7	51.6	38.7	39.9	37.8	32.5	40.8
11	11	16	cuni-a	65.7	49.4	56.4	43.3	60.0	50.3	4.4	19.1	52.3	31.6	30.7	34.8	28.8	39.6
12	14	12	utsamuel	62.8	48.7	54.9	41.6	59.4	49.0	4.3	23.2	53.4	39.5	23.3	34.5	29.7	39.3
13	16	8	unibuc-sb	66.9	36.6	47.3	45.6	46.0	45.8	5.1	22.9	53.5	40.8	31.2	39.8	32.2	39.0
14	12	21	ArgHiTZ	64.7	68.1	66.3	38.0	73.6	50.1	2.9	18.4	48.5	34.9	25.8	32.8	27.2	38.6
15	17	9	KRLabs	57.6	60.6	59.1	34.3	66.4	45.2	5.5	23.4	53.8	38.2	27.8	42.8	31.9	38.6
16	15	19	Loyola	54.6	70.9	61.7	32.8	78.4	46.2	6.2	21.8	50.6	31.5	24.3	34.3	28.1	37.2
17	19	15	aehrc	57.7	41.2	48.1	37.5	49.2	42.5	2.8	20.6	51.3	38.5	28.5	33.4	29.2	35.9
18	18	22	UTSA-NLP	53.2	72.9	61.5	30.4	76.7	43.6	2.5	17.5	51.5	33.9	22.7	30.9	26.5	35.0
19	8	24	DMIS Lab	68.3	62.3	65.2	42.1	70.7	52.8	0.6	12.7	34.8	19.1	9.2	16.2	15.4	34.1
20	20	17	HurLab	56.2	34.8	43.0	36.5	41.6	38.9	4.6	21.0	48.6	37.3	26.5	33.2	28.5	33.7
21	22	11	heiDS	79.0	16.7	27.6	63.0	24.6	35.4	4.4	18.7	51.5	36.1	37.4	33.6	30.3	32.8
22	21	14	DKIT	64.6	23.5	34.5	44.2	29.6	35.5	5.1	21.3	49.8	37.2	27.9	35.1	29.4	32.4
23	23	23	WisPerMed	63.7	27.5	38.4	40.0	31.8	35.4	4.2	18.8	51.7	29.3	24.7	26.0	25.8	30.6
24	24	20	razreshili	40.5	8.1	13.5	30.2	11.1	16.2	2.9	19.4	48.7	32.7	29.4	31.9	27.5	21.9
-	-	-	baseline	83.7	22.9	35.9	65.3	32.8	43.7	2.4	21.0	49.2	39.3	47.0	36.7	32.6	38.1

Table 5: Submission scores using the reconciled answer keys, with clinician-authored answers used for evaluating relevance. Factuality scores are reported at the micro level. O, F, R: Rank using Overall, Factuality (Strict F1), and the average Relevance score. ID: Team identifier; P: Precision; R: Recall; F1: F1 Score; BL: BLEU; RG: ROUGE; SA: SARI; BS: BERTScore; AS: AlignScore; MD: MEDCON; Avg: Overall Relevance Score; OS: Overall Score. All scores are percentages.

systems favored few-shot learning paradigms or the incorporation of synthetic data generated by LLMs over traditional fine-tuning setup. For example, LAMAR (Yoadsanit et al., 2025) created synthetic examples using an LLM to facilitate few-shot prompting with a separate LLM, which was employed to identify relevant note sentences. These sentences were subsequently leveraged to generate the final answer text. In contrast, some systems opted to utilize pre-trained models directly without substantial modifications (e.g., FK).

6 Conclusion

We presented an overview of the ArchEHR-QA Shared Task organized at the BioNLP Workshop in ACL 2025. We discussed the proposed task, the dataset used, the evaluation metrics, and a summary of the baseline and participants' systems. The shared task attracted significant interest, with 29 teams submitting a total of 75 systems and 24 teams providing their system descriptions. Our analysis indicated that systems leveraging proprietary language models achieved higher overall performance, and that top-performing approaches favored few-shot learning strategies over traditional fine-tuning. Additionally, system architectures incorporating an answer reformulation step demonstrated notable improvements in answer quality. The strong interest and competitive submissions underscore the growing momentum in this field. We believe that the insights and resources provided by the ArchEHR-QA Shared Task will promote further advancements in the development and evaluation of EHR-based question answering systems for patient-centered applications.

Rank			Team	Components				Model			Adaptation				Postprocess	
O	F	R	ID	ES →	AG →	CA →	AR	Pty	O-LLM	O-SLM	FT	FS	SD	HT	Mod	Heur
1	1	2	LAMAR	✓	✓			✓				✓	✓			
2	3	1	FK		✓			✓								
3	2	6	unibuc-sd	✓	✓		✓		✓							✓
4	5	5	ssagarwal	✓	✓		✓		✓			✓			✓	
5	10	3	UIC	✓	✓		✓	✓								✓
6	9	4	SzegedAI	✓	✓		✓	✓		✓					✓	
7	6	10	LIMICS		✓	✓	✓	✓							✓	
8	4	18	Neural	✓	✓			✓								
9	7	13	LAILab	✓	✓			✓	✓					✓		
10	13	7	JUNLP		✓			✓				✓				
11	11	16	cuni-a	✓	✓		✓	✓		✓		✓	✓		✓	
12	14	12	utsamuel	✓	✓			✓								
13	16	8	unibuc-sb	✓	✓	✓	✓		✓	✓	✓					✓
14	12	21	ArgHiTZ	✓	✓	✓		✓		✓						✓
15	17	9	KRLabs	✓	✓		✓		✓						✓	
16	15	19	Loyola	✓	✓	✓		✓		✓				✓		
17	19	15	aehrc		✓				✓		✓					
18	18	22	UTSA-NLP	✓	✓				✓			✓				
19	8	24	DMIS Lab	✓	✓			✓								
20	20	17	HurLab	✓	✓				✓	✓	✓					
21	22	11	heiDS	✓	✓				✓							
22	21	14	DKIT	✓	✓				✓	✓	✓					
23	23	23	WisPerMed	✓	✓	✓		✓		✓	✓		✓			
24	24	20	razreshili	✓	✓				✓	✓	✓					

Table 6: Characteristics of the submitted systems with their rankings based on reconciled keys and human answers. O, F, R: Rank using Overall, Factuality (Strict F1), and Relevance score. ID: Team identifier. Broad categories of system components comprise Evidence Selection (ES), Answer Generation (AG), Citation Assignment (CA), and Answer Reformulation (AR). Different types of models employed can be categorized into Proprietary model (Pty), Open-weight large language model (O-LLM), and Open-weight small language model (O-SML). Adaptation or learning methods employed were: Fine Tuning (FT), Few-shot Learning (FS), Use of Synthetic Data (SD), and Hyperparameter Tuning (HT). Postprocessing was performed using the Model itself (Mod) or using Heuristics (Heur).

Limitations

The primary evaluation of system submissions in this shared task relied on automated metrics, which serve as practical proxies for system performance. While such metrics offer scalability and efficiency, they may not fully capture the nuances of answer quality, especially in the clinical domain. Human evaluation that assesses system-generated answers considering the input question and the corresponding clinical note remains the gold standard for determining answer relevance, accuracy, and evidence grounding. However, due to the intensive time and resource requirements, as well as the limited window between the submission deadline and the proceedings release, comprehensive manual evaluation was not feasible within the scope of the shared task. To address this limitation, we plan to conduct a thorough manual assessment of the top submissions from each participating team, focusing on three key criteria: (i) whether the system response adequately answers the question, (ii) whether it leverages relevant clinical evidence, and (iii) whether it uses general knowledge. We anticipate that this forthcoming analysis will provide deeper insights into system performance and help inform and accelerate future development of patient-centered EHR question answering systems.

Acknowledgments

This research was supported by the Division of Intramural Research (DIR) of the National Library of Medicine (NLM), National Institutes of Health, and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

References

Sai Prasanna Teja Reddy Bogireddy, Abrar Majeedi, Viswanath Reddy Gajjala, Zhuoyan Xu, Siddhant Rai, and Vaishnav Potlapalli. 2025. Neural at archehr-qa 2025: Agentic prompt optimization for evidence-grounded clinical question answering. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'Grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H. Kann, Fallon E. Chipidza, Jonathan Leeman, Hugo J. W. L. Aerts, Timothy Miller, Guergana K. Savova, Jack Gallifant, Leo A. Celi, Raymond H. Mak, Maryam Lustberg, Majid Afshar, and Danielle S. Bitterman. 2024. The effect of using a large language model to respond to patient messages. *The Lancet Digital Health*, 0(0).

Adrian Cuadron Cortes, Aimar Sagasti, Maitane Urruela, Iker De la Iglesia, Ane García Domingo-Aldama, Aitziber Atutxa Salazar, Josu Goikoetxea, and Ander Barrena. 2025. Arghitz at archehr-qa 2025: A two-step divide and conquer approach to patient question answering for top factuality. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Patricia Garcia, Stephen P. Ma, Shreya Shah, Margaret Smith, Yejin Jeong, Anna Devon-Sand, Ming Tai-Seale, Kevin Takazawa, Danyelle Clutter, Kyle Vogt, Carlene Lugtu, Matthew Rojo, Steven Lin, Tait Shanafelt, Michael A. Pfeffer, and Christopher Sharp. 2024. Artificial Intelligence–Generated Draft Replies to Patient Inbox Messages. *JAMA Network Open*, 7(3):e243201.

Dragos Dumitru Ghinea and Ștefania Rîncu. 2025. Unibuc-sd at archehr-qa 2025: Prompting our way to clinical qa with multi-model ensembling. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Charlotte J. Haug and Jeffrey M. Drazen. 2023. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. *New England Journal of Medicine*, 388(13):1201–1208.

A. Jay Holmgren, Maria E. Byron, Carrie K. Grouse, and Julia Adler-Milstein. 2023. Association Between Billing Patient Portal Messages as e-Visits and Patient Messaging Volume. *JAMA*, 329(4):339–342.

Hyeon Hwang, Hyeongsoon Hwang, JongMyung Jung, Jaehoon Yun, Minju Song, Yein Park, Dain Kim, Taewhoo Lee, Jiwoong Sohn, Chanwoong Yoon, Sihyeon Park, Jiwoo Lee, Heechul Yang, and Jaewoo Kang. 2025. Dmis lab at archehr-qa 2025: Evidence-grounded answer generation for ehr-based qa via a multi-agent framework. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Tuan Dung Le, Thanh Duong, Shohreh Haddadan, Behzad Jazayeri, Brandon Manley, and Thanh Thieu. 2025. Lailab at archehr-qa 2025: Test-time scaling for evidence selection in grounded question answering from electronic health records. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Siru Liu, Allison B McCoy, Aileen P Wright, Babatunde Carew, Julian Z Genkins, Sean S Huang, Josh F Peterson, Bryan Steitz, and Adam Wright. 2024a. Leveraging large language models for generating responses to patient messages—a subjective analysis. *Journal of the American Medical Informatics Association*, page ocae052.

Siru Liu, Aileen P Wright, Allison B Mccoy, Sean S Huang, Julian Z Genkins, Josh F Peterson, Yaa A Kumah-Crystal, William Martinez, Babatunde Carew, Dara Mize, Bryan Steitz, and Adam Wright. 2024b. Using large language model to guide patients to create efficient and comprehensive clinical care message. *Journal of the American Medical Informatics Association*, 31(8):1665–1670.

Vincent X Liu, Pamela Kaercher, Jennifer Manickam, Eric Smallberg, Kanishka Bhutani, Michelle Mancha, and Kristine Lee. 2024c. Content of Patient Electronic Messages to Physicians in a Large Integrated System. *JAMA Network Open*, 7(4):e244867.

Kathryn A. Martinez, Rebecca Schulte, Michael B. Rothberg, Maria Charmaine Tang, and Elizabeth R. Pfoh. 2024. Patient Portal Message Volume and Time Spent on the EHR: An Observational Study of Primary Care Clinicians. *Journal of General Internal Medicine*, 39(4):566–572.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Yang Ren, Dezhi Wu, Aditya Khurana, George Mastorakos, Sunyang Fu, Nansu Zong, Jungwei Fan, Hongfang Liu, and Ming Huang. 2023. Classification of Patient Portal Messages with BERT-based Language Models. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 176–182.

William R. Small, Batia Wiesenfeld, Beatrix Brandfield-Harvey, Zoe Jonassen, Soumik Mandal, Elizabeth R. Stevens, Vincent J. Major, Erin Lostraglio, Adam Szerencsy, Simon Jones, Yindalon Aphinyanaphongs, Stephen B. Johnson, Oded Nov, and Devin Mann. 2024. Large Language Model–Based Responses to Patients' In-Basket Messages. *JAMA Network Open*, 7(7):e2422399.

Sarvesh Soni and Dina Demner-Fushman. 2025. A Dataset for Addressing Patient's Information Needs related to Clinical Course of Hospitalization. *Preprint*, arXiv:2506.04156.

Mujeen Sung, Song Feng, James Gung, Raphael Shu, Yi Zhang, and Saab Mansour. 2025. Structured List-Grounded Question Answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8347–8359, Abu Dhabi, UAE. Association for Computational Linguistics.

Jui-I Wang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2025. MESAQA: A Dataset for Multi-Span Contextual and Evidence-Grounded Question Answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10891–10901, Abu Dhabi, UAE. Association for Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2023. A survey of consumer health question answering systems. *AI Magazine*, 44(4):482–507.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: A Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. *Scientific Data*, 10(1):586.

Seksan Yoadsanit, Nopporn Lekuthai, Watcharitpol Sermsrisuwan, and Titipat Achakulvisut. 2025. Lamar at archehr-qa 2025: Clinically aligned llm-generated few-shot learning for ehr-grounded patient question answering. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating Factual Consistency with A Unified Alignment Function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Appendix

Tables 7 and 8 provide the factuality scores both at the macro level (averaging per-case F1 scores) and the micro level (aggregating true positives, false positives, and false negatives across all cases).

Rank	Team	Micro						Macro					
		Lenient			Strict			Lenient			Strict		
O	ID	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	DMIS Lab	61.2	59.2	60.2	57.9	59.3	58.6	66.6	67.1	63.2	62.1	69.0	61.2
2	Neural	58.4	63.7	60.9	55.4	63.8	59.3	68.1	69.8	64.8	62.7	71.3	62.6
3	LAILab	59.7	66.0	62.7	56.0	65.5	60.4	67.2	72.1	64.6	62.1	72.8	61.5
4	LAMAR	64.0	53.5	58.3	60.6	53.6	56.9	70.0	62.2	61.8	65.4	64.0	60.2
5	ssagarwal	71.7	35.6	47.6	68.8	36.2	47.5	77.8	44.9	52.1	72.9	46.5	51.4
6	LIMICS	63.6	49.6	55.8	59.9	49.4	54.2	71.0	58.2	59.4	66.6	59.8	57.4
7	cuni-a	60.2	48.1	53.5	56.9	48.1	52.1	66.0	54.2	55.6	61.1	56.0	53.6
8	ArgHiTZ	58.9	65.8	62.1	55.8	65.9	60.5	62.4	69.1	61.9	57.0	69.5	58.5
9	Loyola	51.1	70.5	59.3	48.3	70.5	57.3	56.2	72.9	60.4	52.1	74.0	57.6
10	unibuc-sd	66.5	47.2	55.2	62.7	47.0	53.8	70.8	55.7	58.4	65.7	56.5	56.2
11	SzegedAI	69.7	37.0	48.4	65.6	36.9	47.2	73.6	46.1	53.1	68.3	47.1	51.4
12	KRLabs	50.7	56.6	53.5	48.1	56.8	52.1	60.4	60.6	56.2	55.8	62.3	54.3
13	FK	70.0	37.9	49.2	66.7	38.2	48.6	74.9	49.8	54.5	70.8	51.3	53.4
14	UTSA-NLP	47.0	68.4	55.7	43.7	67.2	53.0	49.6	77.4	56.7	45.1	77.3	52.6
15	UIC	70.4	35.2	46.9	67.3	35.6	46.5	79.1	42.1	51.2	74.7	44.1	51.4
16	utsamuel	55.1	45.3	49.7	51.4	44.7	47.8	57.0	55.4	51.8	52.2	56.0	49.0
17	aehrc	55.5	42.0	47.8	52.9	42.4	47.1	65.4	48.0	50.4	61.4	49.5	49.1
18	unibuc-sb	61.7	35.9	45.4	58.7	36.1	44.7	68.5	41.4	47.8	63.6	42.7	46.4
19	HurLab	52.9	34.8	41.9	49.3	34.3	40.4	61.2	42.0	44.8	56.7	43.0	42.7
20	JUNLP	57.5	32.4	41.4	54.2	32.3	40.5	62.4	43.6	46.9	58.4	45.0	45.8
21	WisPerMed	59.1	27.1	37.1	55.4	26.9	36.2	59.5	33.9	39.9	54.0	34.0	37.7
22	DKIT	59.9	23.1	33.4	56.5	23.1	32.7	63.4	31.1	36.5	60.0	32.4	35.9
23	heiDS	71.2	16.0	26.2	67.7	16.1	26.0	73.9	22.5	30.7	69.7	24.0	30.7
24	razreshili	39.7	8.4	13.9	36.8	8.2	13.5	53.8	13.6	19.1	49.6	14.5	19.0
-	baseline	77.0	22.3	34.6	71.6	21.9	33.6	83.0	30.8	39.9	77.4	31.5	39.0

Table 7: Factuality scores using the preliminary version of answer keys, with both micro and macro level calculations. *O*: Rank using Overall score. ID: Team identifier; P: Precision; R: Recall; F1: F1 Score. All scores are percentages.

Rank	Team	Micro						Macro					
		Lenient			Strict			Lenient			Strict		
O	ID	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	LAMAR	72.9	57.4	64.2	49.6	72.0	58.8	78.3	67.2	68.4	56.7	75.5	61.9
2	FK	78.8	40.3	53.3	59.1	55.6	57.3	81.8	52.5	58.7	64.3	60.8	59.8
3	unibuc-sd	75.4	50.4	60.4	53.0	65.2	58.4	80.2	59.8	65.2	60.7	69.1	62.3
4	ssagarwal	79.0	37.0	50.4	58.3	50.3	54.0	84.6	47.5	55.9	67.0	55.2	56.8
5	UIC	77.6	36.6	49.7	55.5	48.2	51.6	86.8	43.7	54.1	68.7	51.1	54.5
6	SzegedAI	78.4	39.3	52.3	54.4	50.2	52.2	80.7	47.9	56.2	58.5	53.8	53.4
7	LIMICS	71.2	52.3	60.3	46.9	63.5	54.0	78.6	61.5	64.4	55.5	67.2	57.5
8	Neural	67.5	69.3	68.4	42.8	81.0	56.0	76.5	75.6	72.3	54.6	82.8	62.1
9	LAILab	65.7	68.4	67.0	40.6	77.8	53.3	73.3	74.3	68.9	51.9	79.9	57.8
10	JUNLP	64.9	34.5	45.0	49.5	48.5	49.0	68.9	46.5	51.0	53.6	54.4	52.0
11	cuni-a	65.7	49.4	56.4	43.3	60.0	50.3	72.5	56.6	59.3	54.5	64.3	54.8
12	utsamuel	62.8	48.7	54.9	41.6	59.4	49.0	63.8	58.0	56.6	45.1	63.8	50.6
13	unibuc-sb	66.9	36.6	47.3	45.6	46.0	45.8	72.5	43.2	49.5	54.1	48.9	48.4
14	ArgHiTZ	64.7	68.1	66.3	38.0	73.6	50.1	68.4	71.1	65.6	45.8	74.4	52.8
15	KRLabs	57.6	60.6	59.1	34.3	66.4	45.2	67.4	64.9	62.0	48.0	69.0	51.9
16	Loyola	54.6	70.9	61.7	32.8	78.4	46.2	59.0	73.9	62.4	40.1	78.4	49.9
17	aehrc	57.7	41.2	48.1	37.5	49.2	42.5	68.7	47.6	51.4	51.5	52.7	47.8
18	UTSA-NLP	53.2	72.9	61.5	30.4	76.7	43.6	56.4	80.3	61.7	35.5	79.8	46.2
19	DMIS Lab	68.3	62.3	65.2	42.1	70.7	52.8	73.9	70.3	68.6	50.5	74.7	57.2
20	HurLab	56.2	34.8	43.0	36.5	41.6	38.9	63.9	41.3	45.4	45.6	44.5	41.1
21	heiDS	79.0	16.7	27.6	63.0	24.6	35.4	82.4	24.5	32.9	66.3	28.5	36.7
22	DKIT	64.6	23.5	34.5	44.2	29.6	35.5	68.1	32.3	38.3	51.1	34.6	37.6
23	WisPerMed	63.7	27.5	38.4	40.0	31.8	35.4	62.9	34.8	40.8	39.7	35.1	35.3
24	razreshili	40.5	8.1	13.5	30.2	11.1	16.2	57.8	13.4	18.3	47.4	14.6	20.2
-	baseline	83.7	22.9	35.9	65.3	32.8	43.7	89.2	32.4	42.2	74.2	38.6	47.3

Table 8: Factuality scores using the reconciled answer keys, with both micro and macro level calculations. *O*: Rank using Overall score. ID: Team identifier; P: Precision; R: Recall; F1: F1 Score. All scores are percentages.