Overview of the ArchEHR-QA 2026 Shared Task on Grounded Question Answering from Electronic Health Records

Sarvesh Soni, 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

We present an overview of the ArchEHR-QA 2026 Shared Task on grounded question answering from electronic health records (EHRs), organized at the CL4Health Workshop at LREC 2026. The 2026 task decomposes grounded EHR question answering (QA) into four complementary subtasks: question interpretation, evidence identification, answer generation, and evidence alignment. We evaluated submitted systems for the text-generation subtasks (question interpretation and answer generation) using lexical, semantic, and grounding-sensitive automatic metrics, and for the evidence-centric subtasks (evidence identification and evidence alignment) using precision, recall, and F1. The shared task received 198 submitted runs from 43 teams, and 17 teams additionally provided system descriptions for this overview. The highest-ranked systems differed across subtasks, and gains over the organizer baseline were largest on the evidence-centric subtasks. Across submitted system descriptions, prompt-based large language model (LLM) pipelines were dominant, whereas task-specific fine-tuning was rare; retrieval, self-consistency, and ensembling were especially common in the strongest evidence-centric systems. In this paper, we describe the task design, data, evaluation protocol, baselines, participation, official results, and common system characteristics, and discuss implications for developing clinically faithful and transparent QA systems.

Keywords: grounded question answering, electronic health records, clinical natural language processing, question reformulation, evidence attribution, shared task evaluation

1. Introduction

Patients frequently contact their healthcare providers with questions about diagnoses, treatments, test results, and events documented in their electronic health records (EHRs) (Tai-Seale et al., 2017). Patient messaging is also a substantial contributor to clinician burden (Yan et al., 2021; Martinez et al., 2024). From a clinical natural language processing (NLP) perspective, responding to a patient question from the EHR can be decomposed into four related steps: identifying the patient's core information need, locating answer-bearing clinical evidence, drafting an answer, and linking the answer back to the supporting evidence. Systems that assist with any subset of these steps could reduce clinician burden while improving transparency.

Clinicians often find it challenging and time-consuming to process patient messages that are long, verbose, and emotionally charged (Lanham et al., 2018). Patients often ask longer questions, and as a result, prior work has studied summarization of consumer health questions that can be answered from general medical knowledge (Ben Abacha et al., 2021). However, comparatively less work has focused on reformulating patient-authored narratives into patient-specific clinical questions that can then be answered from the patient's EHR (Soni and Demner-Fushman, 2025b).

EHR-grounded response drafting also requires effective evidence identification. Because EHRs contain large amounts of heterogeneous patient information, finding the specific note sentences needed to answer a question is time-consuming (Koopman et al., 2015; Nijor et al., 2022). Although clinical information retrieval has been studied extensively, much of that literature concerns retrieval across patients or documents rather than extraction of patient-specific answer evidence within an individual patient's EHR (Sivarajkumar et al., 2024).

Finally, clinicians prefer QA systems whose answers can be verified against clinical evidence (Kell et al., 2024). Grounding (linking text back to the source evidence that supports it) is therefore a central requirement for trustworthy clinical QA (Chandu et al., 2021). In the present setting, grounding should allow a clinician to inspect which parts of the EHR support each answer sentence, regardless of whether the draft answer was generated by a model, authored by a clinician, or jointly produced.

To study these interdependent capabilities, we organized the ArchEHR-QA 2026 Shared Task[1] as part of the CL4Health Workshop at LREC 2026. This is the second ArchEHR-QA shared task. The 2025 edition, organized at the BioNLP Workshop at ACL 2025, focused on end-to-end answer generation with citations to supporting note sentences (Soni et al., 2025). In contrast, the 2026 edition decomposes grounded EHR QA into four complemen-

[1] archehr-qa.github.io

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 255–267
12 May 2026.

tary subtasks: question interpretation, evidence identification, answer generation, and evidence alignment. This decomposition permits a more fine-grained evaluation of the capabilities required for grounded response drafting. In this paper, we describe the task design, data, evaluation protocol, baseline, participation, official results, and observations from the submitted systems.

2. ArchEHR-QA 2026 Task Description

ArchEHR-QA 2026 consists of four subtasks: Question Interpretation, Evidence Identification, Answer Generation, and Evidence Alignment. Table 1 presents a running example illustrating the inputs and reference outputs for all four subtasks.

2.1. Subtask 1: Question Interpretation

Subtask 1 frames question interpretation as a constrained reformulation task. Given a free-text patient-authored question, the system must produce a concise clinician-interpreted question that captures the patient's core clinical information need. The target output is a single well-formed question phrased as the type of query a clinician would submit to an intelligent EHR system when preparing a response. To preserve the focus on concise reformulation, outputs were limited to 15 words and were expected to avoid unsupported clinical additions. In the running example, the reference output is the clinician-interpreted question shown in Table 1. The clinician-interpreted question provides a concise reformulation of the patient's core information need. For the downstream subtasks, however, we provided both the original patient-authored question and the clinician-interpreted question, since the two forms may contain complementary information, and participants could draw on either or both.

Given the patient question, the clinician-interpreted question, clinical specialty labels, and a sentence-segmented clinical note excerpt, the system must identify the note sentences that provide the evidence needed to answer the question.

2.2. Subtask 2: Evidence Identification

Subtask 2 frames evidence identification as sentence-level evidence selection. Given the patient question, the clinician-interpreted question, clinical specialty labels, and a sentence-segmented clinical note excerpt, the system must identify the note sentences that provide the evidence needed to answer the question. The target output is a set of sentence IDs corresponding to evidence that is sufficient for answering the question. This subtask emphasizes identification of essential evidence while discouraging unnecessary or only tangentially related sentences. The reference evidence set for the running example is shown in Table 1.

2.3. Subtask 3: Answer Generation

Subtask 3 frames answer generation as grounded response drafting. Given the patient question, the clinician-interpreted question, clinical specialty labels, and the clinical note excerpt, the system must generate a concise answer supported by the note. Answers were limited to 75 words, roughly corresponding to a short paragraph of up to five sentences. We selected this limit based on preliminary baseline experiments and prior work suggesting that paragraph-length answers are generally preferred by users (Lin et al., 2003; Jeon et al., 2006). Reference answers were written in a professional clinical register and grounded in the note content, without unsupported speculation. Unlike Subtask 2, this subtask does not provide identified evidence sentences as input; instead, it evaluates answer generation directly from the note excerpt rather than from a pre-filtered evidence set. This keeps the task focused on grounded answer generation, which still requires locating relevant information within the note, while allowing teams to use evidence identified by their own systems, including predictions from Subtask 2. The reference answer for the running example is shown in Table 1.

2.4. Subtask 4: Evidence Alignment

Subtask 4 frames grounding as sentence-level answer-to-evidence alignment. Given the question context, clinical specialty labels, a sentence-numbered clinical note excerpt, and an answer to be grounded, the system must predict which note sentence(s) support each answer sentence. The target output is therefore a set of mappings from answer sentences to note sentence IDs. These alignments are many-to-many: one answer sentence may be supported by multiple note sentences, and one note sentence may support multiple answer sentences. Answer sentences without direct support may be assigned an empty evidence set. For the shared task, we intentionally used the clinician-authored reference answer as the input answer for this subtask. This design isolates the alignment problem from answer-generation errors and enables straightforward automatic evaluation against gold answer-to-evidence mappings. It also provides a controlled setting for developing methods that may later be extended to system-generated answers in end-to-end pipelines. In the running example, the gold output is the set of sentence-level answer-to-evidence mappings shown in Table 1.

ArchEHR-QA Patient Case

PQ **Patient question:** I just wrote about my dad given multiple shots of lasciks after he was already so swelled his shin looked like it would burst open. Why would they give him so much. He was on oxygen and they took him off of the higher flow rate.

CQ **Clinician-interpreted question:** Why was he given lasix and his oxygen flow rate was reduced?

SP **Clinical specialty label(s):** Cardiology

NOTE **Clinical note excerpt:**
1: Acute diastolic heart failure: Pt developed signs and symptoms of volume overload with shortness of breath, increased oxygen requirement and lower extremity edema. **2:** Echo showed preserved EF, no WMA and worsening AI. **3:** CHF most likely secondary to worsening valvular disease. **4:** He was diuresed with lasix IV, intermittently on lasix gtt then transitioned to PO torsemide with improvement in symptoms, although remained on a small amount of supplemental oxygen for comfort.

5: Respiratory failure: The patient was intubated for lethargy and acidosis initially and was given 8 L on his presentation to help maintain his BP's. **6:** This undoubtedly contributed to his continued hypoxemic respiratory failure. **7:** He was advanced to pressure support with stable ventilation and oxygenation. **8:** On transfer to the CCU patient was still intubated but off pressors. **9:** Patient was extubated successfully. **10:** He was reintubated transiently for 48 hours for urgent TEE and subsequently extubated without adverse effect or complication.

REL **Note Sentence Relevance Annotations:** `[1, 4, 5, 6]` → *essential*; `[3, 7]` → *supplementary*

ANS **Reference answer (clinician-authored): A1:** The patient was given Lasix for acute diastolic heart failure with symptoms including shortness of breath and lower extremity edema. **A2:** The patient was given 8 liters of fluid to help maintain his blood pressure which contributed to his respiratory failure. **A3:** After the patient's heart failure was treated with Lasix, he showed improvement in shortness of breath and his oxygen requirement, and he only needed to remain on a small amount of oxygen for comfort.

ALN **Answer to Note Sentence Alignments: A1** → `[1, 4]` **A2** → `[5, 6]` **A3** → `[4]`

ST1: Question interpretation

Inputs: PQ *Patient question*. **Target output:** CQ *Clinician-interpreted question.*

Sample system response (good): Why were multiple doses of Lasix given despite decreased oxygen requirements?
Sample system response (bad – generic/malformed): Lasix dosage protocol for severe edema with oxygen therapy?

ST2: Evidence identification

Inputs: PQ *Patient question*; CQ *clinician-interpreted question*; SP *clinical specialty label(s)*; NOTE *clinical note*.
Target output: REL *Note Sentence Relevance Annotations*

Sample system response (missing key evidence; P 1.00, R 0.50, F1 0.67): `[1, 4]`
Sample system response (over-inclusive; P 0.50, R 1.00, F1 0.67): `[1, 4, 5, 6, 7, 8, 9, 10]`

ST3: Answer generation

Inputs: PQ *Patient question*; CQ *clinician-interpreted question*; SP *clinical specialty label(s)*; NOTE *clinical note*.
Target output: ANS *Reference answer.*

Sample system response (bad – generic/not grounded):
Lasix is commonly used to remove excess fluid when patients are swollen or having trouble breathing, and oxygen can often be reduced once breathing improves. Changes in oxygen flow usually reflect better respiratory status.
Sample system response (ok – some grounding but mostly generic):
He had fluid overload from heart failure, so Lasix was used to reduce swelling and shortness of breath. It also reports improvement after diuresis, with only a small amount of supplemental oxygen needed for comfort.

ST4: Evidence alignment

Inputs: PQ *Patient question*; CQ *clinician question*; SP *specialty*; NOTE *clinical note*; ANS *reference answer*.
Target output: ALN *Answer to Note Sentence Alignments*

Sample response (under-citing; P 0.75, R 0.60, F1 0.67): **A1:** → `[1].` **A2:** → `[5].` **A3:** → `[4,5].`
Sample response (over-citing; P 0.50, R 1.00, F1 0.67): **A1:** → `[1,4,5].` **A2:** → `[5,6,7].` **A3:** → `[1,4,6,7].`

Table 1: Example showing an ArchEHR-QA patient case and the corresponding inputs and target outputs for all four subtasks, with representative sample system responses.

	Dev (n=20)	**Test** (n=100)	**Test-2026** (n=47)
Mean word count			
Patient question	85.2	92.3	94.1
Clinician question	10.8	10.6	10.4
Note excerpt	320.8	380.4	497.2
Clinician answer	73.6	72.4	72.7
Mean note sentences			
Total	21.4	26.0	34.0
Essential	6.0 (28.3%)	6.6 (25.3%)	9.7 (28.6%)
Supplementary	1.3 (6.1%)	5.5 (21.3%)	6.9 (20.4%)
Not relevant	14.1 (65.7%)	13.9 (53.4%)	17.3 (51.0%)

Table 2: Dataset statistics by split. Values are mean word counts for text fields and mean sentence counts for note relevance categories; percentages denote proportions of note sentences within each split. Test-2026 is the official test split for Subtasks 1–3; Subtask 4 uses both Test and Test-2026.

3. Data Description

ArchEHR-QA 2026 builds on the ArchEHR-QA dataset introduced in prior work (Soni and Demner-Fushman, 2026). That work describes the dataset creation process, including the annotation schema and guidelines. The dataset contains patient cases drawn from intensive care and emergency department settings with hospitalization-related questions. Table 1 shows one illustrative example. The dataset pairs patient-authored questions, inspired by real patient information needs expressed in public health forums, with sentence-segmented clinical note excerpts derived from the MIMIC database. Each data instance is referred to as a patient *case*. Each case contains: (i) a free-text patient question; (ii) a clinician-interpreted question; (iii) a sentence-segmented clinical note excerpt; (iv) sentence-level relevance annotations for the note sentences (*essential*, *supplementary*, or *not relevant*); (v) a clinician-authored reference answer; (vi) sentence-level answer-to-evidence alignments; and (vii) one or more clinical specialty labels.

ArchEHR-QA 2026 uses 167 patient cases. Of these, 134 cases come from the original release (Soni and Demner-Fushman, 2026), and 33 cases were newly curated for the 2026 shared task using the same annotation schema.

3.1. Official Evaluation Splits and Staged Release Schedule

The official test set for Subtasks 1–3 comprised case IDs 121–167 (47 cases), while case IDs 1–120 were released for system development. We restricted official evaluation of Subtasks 1–3 to case IDs 121–167 because case IDs 1–120 were already publicly used in the 2025 iteration of ArchEHR-QA (Soni et al., 2025). For Subtask 1, the prior release of clinician-interpreted questions made those cases unsuitable for a new evaluation. We used the same held-out split for the closely related Subtasks 2–3 to avoid reusing cases that had already been used in a similar shared task setting in 2025. By contrast, Subtask 4 was newly introduced in 2026 and had not previously been evaluated on those cases. We therefore evaluated Subtask 4 on the larger split of case IDs 21–167 (147 cases), reserving case IDs 1–20 for development. Table 2 summarizes dataset statistics across the splits.

The official evaluation followed a staged release schedule designed to mirror the intended workflow and preserve separation between subtasks. For Subtask 1, participants initially received only the patient-authored questions for the 47 held-out test cases. After the Subtask 1 deadline, the release for Subtasks 2–3 added the clinician-interpreted questions and clinical note excerpts for the same 47 cases. After the Subtasks 2–3 deadline, the release for Subtask 4 provided case IDs 21–167 together with sentence-numbered reference answers to be aligned to the corresponding note excerpts. Sentence-level relevance annotations were withheld from all official test releases.

4. Evaluation

4.1. Metrics

Each subtask was evaluated independently on its corresponding held-out test set. Official leaderboard rankings were determined by a single primary metric per subtask, while additional metrics were reported to provide a broader characterization of system behavior. The scoring scripts are available on GitHub[2].

Subtask 1. The system-generated clinician-interpreted questions were compared with the reference questions using ROUGE-Lsum (Lin, 2004),

[2] github.com/soni-sarvesh/archehr-qa-2026

BERTScore (Zhang et al., 2019), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023). These metrics capture complementary aspects of reformulation quality: ROUGE-Lsum measures lexical overlap, BERTScore measures semantic similarity, AlignScore measures factual consistency, and MEDCON measures clinical concept overlap. To enforce the task constraint, outputs longer than 15 words were truncated to the first 15 words before scoring. The official ranking metric was the Overall Score (OS), computed as the unweighted arithmetic mean of the four metric values.

Subtask 2. The predicted evidence sentence IDs were compared with the reference evidence annotations using Precision, Recall, and F1. We report both micro-averaged scores, which pool predictions across all cases, and macro-averaged scores, which average case-level scores. Results are reported under two evaluation variants. Under *Strict* evaluation, only sentences labeled *essential* are treated as gold evidence. Under *Lenient* evaluation, predicted *supplementary* sentences were not counted as false positives. The official ranking metric was Strict Micro F1.

Subtask 3. The system-generated answers were compared with the reference answers using BLEU (Papineni et al., 2002), ROUGE-Lsum, SARI (Xu et al., 2016), BERTScore, AlignScore, and MEDCON. Together, these metrics capture complementary aspects of answer quality, with BLEU and ROUGE-Lsum emphasizing lexical overlap, SARI capturing content-editing behavior, and BERTScore, AlignScore, and MEDCON providing semantic, factual, and clinical-concept perspectives. Prior work in this setting suggests that automatic metrics can provide useful system-ranking signals when computed against clinician-authored reference answers (Soni and Demner-Fushman, 2025a). The official ranking metric was the Overall Score (OS), computed as the unweighted arithmetic mean of the six metric values.

Subtask 4. The predicted alignments were evaluated as sets of links between answer sentences and note sentences, where each link has the form (*answer sentence* k $\rightarrow$ *note sentence* i). We report Precision, Recall, and F1 at both the micro and macro levels. Micro-averaged scores pool predicted and reference links across all cases, whereas macro-averaged scores average case-level scores. Over-citing is penalized through false positives, which lowers Precision and therefore F1. The official ranking metric was Micro F1.

4.2. Baseline

We used a zero-shot baseline based on Qwen/Qwen3-4B-Instruct-2507 (Yang et al., 2025) for all four subtasks. No task-specific fine-tuning or few-shot exemplars were used. The model was prompted in a single-turn chat format with task-specific instructions; the prompt templates are provided in Appendix Tables 8–11. We used the same decoding settings across subtasks, following the Qwen3-recommended hyperparameters[3]: temperature $= 0.7$, top-$p = 0.8$, top-$k = 20$, and min-$p = 0.0$.

For Subtask 1, the model rewrote each patient-authored question as a clinician-interpreted question (`max_new_tokens=256`). For Subtask 2, the model predicted supporting sentence IDs from the clinical note excerpt conditioned on the patient-authored question, the clinician-interpreted question, and the specialty label (`max_new_tokens=512`). For Subtask 3, the model generated a clinically grounded answer from the patient-authored question, the clinician-interpreted question, the specialty label, and the clinical note excerpt (`max_new_tokens=512`). For Subtask 4, the model predicted answer-to-evidence sentence alignments from the clinician-interpreted question, the sentence-numbered clinical note excerpt, and the sentence-numbered answer (`max_new_tokens=1024`). Although both question forms were available for Subtask 4, the baseline used only the clinician-interpreted question, based on preliminary experiments on the development set.

5. Participation

We used Codabench[4] to collect system predictions from teams (Xu et al., 2022). Teams could participate in any subset of the four subtasks and were allowed to submit up to three runs per subtask. For the overview paper, we asked each team to nominate its best run per subtask and to submit a short system description.

5.1. Participating Teams

We received 198 submitted runs from 43 teams across the four subtasks: 41 runs for Subtask 1, 59 for Subtask 2, 35 for Subtask 3, and 63 for Subtask 4. Among these 43 teams, 17 submitted system descriptions and are therefore included in the

[3] huggingface.co/Qwen/Qwen3-4B-Instruct-2507

[4] Subtask 1: codabench.org/competitions/12865; Subtask 2: codabench.org/competitions/13526; Subtask 3: codabench.org/competitions/13527; Subtask 4: codabench.org/competitions/13528

Team ID	Affiliation	Country	ST1	ST2	ST3	ST4
BIT.UA-AAUBS	University of Aveiro; Aalborg University	Portugal; Denmark	✓	✓	✓	✓
CaresAI	Cairo University; Techvify	Egypt; Vietnam	✓	✓	×	×
GigitAI	GigitAI	USA	×	✓	✓	×
HealthNLP_Retrievers	University of Maryland Baltimore County; Jahangirnagar University	USA; Bangladesh	✓	✓	✓	✓
HiTZ-IXA	University of the Basque Country	Spain	×	✓	×	✓
KA	L&T Technology Services	India	×	×	×	✓
KPSCMI	Kaiser Permanente Southern California	USA	✓	✓	×	✓
MedEvi-NS	University of Aberdeen; Bournemouth University	UK	×	×	×	✓
Neural	University of Chicago; University of Wisconsin Madison; Lovely Professional University	USA; India	✓	✓	✓	✓
OptiMed	Queen Mary University of London; Independent Researcher	UK; Turkey	✓	✓	✓	✓
razreshili	Independent Researcher	Germany	×	×	×	✓
sebis	Technical University of Munich	Germany	✓	✓	✓	✓
TAMU-NLP-Lab	Texas A&M University	USA	×	✓	✓	×
tt501	Vietnam National University	Vietnam	×	✓	✓	✓
UIC-AIHealth4All	University of Illinois Chicago	USA	×	✓	✓	✓
WisPerMed	University of Applied Sciences and Arts Dortmund; University of Duisburg-Essen	Germany	✓	✓	✓	✓
Yale-DM-Lab	Yale University; UMass Lowell	USA	✓	✓	✓	✓

Table 3: Participating teams who submitted a system description, their affiliation, country, and the subtasks they participated in. Rows are sorted by Team ID in alphabetical order. ST1–ST4 denote Subtasks 1–4.

analysis below. Table 3 lists those teams, their affiliations, countries, and the subtasks they entered. Among the teams represented in the overview paper, seven participated in all four subtasks.

5.2. Results

Tables 4–7 report the official results for subtasks 1–4. All highest-ranked submissions outperformed the organizer baseline, but the margin varied substantially across subtasks: +11.3 Overall Score for Subtask 1 (31.2 vs. 19.9), +14.1 Strict Micro F1 for Subtask 2 (63.7 vs. 49.6), +4.3 Overall Score for Subtask 3 (36.3 vs. 32.0), and +16.2 Micro F1 for Subtask 4 (81.5 vs. 65.3). The baseline remained competitive with lower-ranked submissions in Subtasks 1 and 3, whereas the evidence-centric subtasks showed a clearer separation between the strongest submitted systems and the zero-shot baseline.

For Subtask 1 (Question Interpretation; Table 4), the highest-ranked system was submitted by *HealthNLP_Retrievers* and obtained an overall score of 31.2, followed by *KPSCMI* (30.8) and *OptiMed* (29.9). The top three systems were separated by only 1.3 overall-score points. The strongest systems emphasized different aspects of the task: the *HealthNLP_Retrievers* submission obtained the highest ROUGE-Lsum and BERTScore,

⋆	Team	Relevance				
		RG	BS	AS	MD	OS
1	HealthNLP_Retrievers	35.3	46.8	24.0	18.7	31.2
2	KPSCMI	27.8	41.0	26.4	27.9	30.8
3	OptiMed	28.8	43.1	27.7	19.9	29.9
4	Neural	31.3	43.6	15.2	25.6	28.9
5	Yale-DM-Lab	28.2	40.6	19.7	19.8	27.1
6	WisPerMed	21.8	38.0	21.7	26.3	26.9
7	sebis	22.9	36.9	21.4	21.3	25.6
8	CaresAI	22.5	36.1	12.5	24.5	23.9
9	BIT.UA-AAUBS	17.9	29.4	8.7	20.2	19.0
-	baseline	23.4	31.0	7.9	17.3	19.9

Table 4: Subtask 1 (Question Interpretation) results. Rows are ranked (⋆) by the overall score (OS). RG: ROUGE-Lsum; BS: BERTScore; AS: AlignScore; MD: MEDCON; OS: Overall score.

Rank	Team	Micro						Macro					
		Strict			Lenient			Strict			Lenient		
		P	R	F1†	P	R	F1	P	R	F1	P	R	F1
1	Neural	60.2	67.6	63.7	77.8	67.6	72.4	64.3	69.7	64.8	81.9	69.7	73.1
2	OptiMed	56.7	71.3	63.2	73.1	71.3	72.2	61.5	73.0	64.1	78.1	73.0	72.9
3	UIC-AIHealth4All	59.3	67.0	62.9	74.3	67.0	70.4	61.8	69.2	63.0	77.8	69.2	70.5
4	Yale-DM-Lab	52.3	75.7	61.9	68.0	75.7	71.6	56.7	75.5	61.1	73.8	75.5	70.4
5	TAMU-NLP-Lab	56.7	65.9	60.9	76.0	65.9	70.6	63.0	66.7	59.9	80.0	66.7	69.1
6	HealthNLP_Retrievers	58.4	62.1	60.2	75.7	62.1	68.3	65.7	62.5	59.9	79.8	62.5	67.1
7	KPSCMI	46.5	81.8	59.3	62.4	81.8	70.8	51.9	82.2	60.9	69.4	82.2	72.0
8	tt501	56.0	61.9	58.8	74.9	61.9	67.8	61.3	63.9	58.4	78.4	63.9	66.8
9	WisPerMed	67.1	52.3	58.8	84.2	52.3	64.5	69.8	53.4	58.1	84.9	53.4	63.6
10	BIT.UA-AAUBS	51.0	69.4	58.8	68.5	69.4	68.9	59.8	69.1	59.3	76.9	69.1	68.9
11	GigitAI	46.5	77.0	58.0	62.1	77.0	68.8	52.8	77.7	59.1	69.6	77.7	69.3
12	CaresAI	46.7	67.6	55.3	61.3	67.6	64.3	50.1	68.2	55.2	66.8	68.2	64.0
13	sebis	37.9	80.7	51.6	49.1	80.7	61.1	40.8	80.3	52.6	54.1	80.3	62.6
14	HiTZ-IXA	44.2	47.9	46.0	56.3	47.9	51.8	48.7	48.7	46.7	62.4	48.7	52.2
-	baseline	51.0	48.1	49.6	64.5	48.1	55.1	60.7	47.2	47.9	74.8	47.2	52.8

Table 5: Subtask 2 (Evidence Identification) results. Rows are ranked by Strict Micro F1 Score (†). P: Precision; R: Recall; F1: F1 score.

Rank	Team	Lexical			Semantic			Overall
		BL	RG	SA	BS	AS	MD	OS
1	WisPerMed	9.9	27.8	58.6	46.8	31.7	43.1	36.3
2	TAMU-NLP-Lab	9.7	27.4	59.6	46.2	34.5	39.9	36.2
3	BIT.UA-AAUBS	8.6	26.4	60.0	45.0	30.2	43.2	35.6
4	Neural	9.4	25.6	57.7	43.5	34.3	40.9	35.2
5	HealthNLP_Retrievers	7.0	25.4	59.2	43.8	33.6	38.7	34.6
6	OptiMed	5.7	25.2	56.5	43.1	37.4	39.1	34.5
7	UIC-AIHealth4All	6.1	24.2	57.2	41.8	24.5	37.5	31.9
8	GigitAI	5.1	21.1	56.8	40.0	28.4	39.4	31.8
9	sebis	3.8	22.0	56.6	39.3	33.9	33.4	31.5
10	tt501	6.2	24.0	58.3	41.4	22.4	35.9	31.4
11	Yale-DM-Lab	9.1	23.1	56.5	37.2	22.1	37.6	31.0
-	baseline	5.0	22.5	56.2	40.6	32.2	35.4	32.0

Table 6: Subtask 3 (Answer Generation) results. Rows are ranked by the overall score (OS). BL: BLEU; RG: ROUGE-Lsum; SA: SARI; BS: BERTScore; AS: AlignScore; MD: MEDCON; OS: Overall score.

the *OptiMed* submission obtained the highest AlignScore, and the *KPSCMI* submission obtained the highest MEDCON.

In Subtask 2 (Evidence Identification; Table 5), the *Neural* system achieved the best Strict Micro F1 (63.7), narrowly ahead of *OptiMed* (63.2) and *UIC-AIHealth4All* (62.9). Lenient evaluation improved F1 for all systems, indicating that many submissions selected supplementary but still clinically plausible evidence sentences. The systems showed different precision–recall profiles: *KPSCMI* and *sebis* favored higher recall, whereas *Neural* and *OptiMed* achieved a more balanced trade-off.

For Subtask 3 (Answer Generation; Table 6), the *WisPerMed* system obtained the highest overall score (36.3), with *TAMU-NLP-Lab* effectively tied at 36.2 and *BIT.UA-AAUBS* third at 35.6. Metric leaders again differed by evaluation perspective: the *WisPerMed* submission obtained the best BLEU, ROUGE-Lsum, and BERTScore; the *BIT.UA-AAUBS* submission obtained the best SARI and MEDCON; and the *OptiMed* submission obtained the best AlignScore. This result suggests that answer generation systems are ranked differently depending on whether the evaluation emphasizes lexical overlap, semantic similarity, or clinical

Rank	Team	Micro			Macro		
		P	R	F1†	P	R	F1
1	BIT.UA-AAUBS	88.0	75.9	81.5	88.3	78.9	82.2
2	WisPerMed	86.9	76.3	81.3	87.5	79.4	82.1
3	Yale-DM-Lab	83.3	77.7	80.4	84.8	80.7	81.8
4	OptiMed	80.7	79.8	80.3	82.8	82.8	81.6
5	UIC-AIHealth4All	83.6	76.3	79.8	85.5	79.5	80.9
6	tt501	78.2	80.1	79.1	80.0	82.6	79.7
7	Neural	84.3	73.7	78.6	85.8	76.6	79.4
8	KPSCMI	86.2	71.5	78.1	86.4	75.5	78.9
9	HealthNLP_Retrievers	83.8	71.1	76.9	85.3	74.1	77.4
10	MedEvi-NS	75.7	77.4	76.5	79.3	80.8	78.3
11	sebis	79.6	70.6	74.8	82.6	73.6	76.2
12	razreshili	73.8	63.0	67.9	75.8	67.6	69.5
13	HiTZ-IXA	74.7	60.3	66.8	73.5	65.4	66.7
14	KA	68.1	64.5	66.2	70.9	69.0	68.0
-	baseline	59.1	72.9	65.3	72.0	75.3	71.1

Table 7: Subtask 4 (Evidence Alignment) results. Rows are ranked by Micro F1 Score (†). P: Precision; R: Recall; F1: F1 score.

concept preservation. Compared with the other subtasks, Subtask 3 showed a relatively small margin over the organizer baseline (36.3 vs. 32.0).

For Subtask 4 (Evidence Alignment; Table 7), the highest-ranked system was submitted by *BIT.UA-AAUBS* and obtained a Micro F1 of 81.5, followed closely by *WisPerMed* (81.3), *Yale-DM-Lab* (80.4), and *OptiMed* (80.3). The top four systems were separated by only 1.2 F1 points. Precision was generally high among the strongest submissions, while Recall accounted for most of the remaining variation.

Among teams participating in all four subtasks, *OptiMed* showed the most consistently strong performance, placing within the top six in every subtask. No single team submitted the highest-ranked system in more than one subtask.

5.3. Approaches

Across all four subtasks, prompt-based use of large language models (LLMs) was dominant. Task-specific fine-tuning was comparatively rare; more common strategies were prompt engineering, retrieval of similar development examples, repeated sampling with self-consistency, and multi-model ensembling. Strong submissions were obtained with both API-based models and locally hosted or open-weight models, suggesting that no single deployment setup dominated the task.

For Subtask 1 (Question Interpretation), in-context prompting with examples was frequent: 8/9 systems used few-shot or many-shot prompting, often with explicit post-processing to enforce the single-question output format and the 15-word limit. Five of the nine systems used proprietary API-based LLMs, whereas the remainder relied on locally hosted models. Three systems introduced an additional reasoning or decomposition step before generation, e.g., by extracting important medical terms or by scoring candidate questions from multiple models. Prompt optimization was reported by two systems, using frameworks such as MIPROv2 (Opsahl-Ong et al., 2024) or GEPA (Agrawal et al., 2025). The strongest systems on this subtask followed three different designs: few-shot prompting with tight output constraints on an API-based model (*HealthNLP_Retrievers*), many-shot prompting with the full development set as examples on a locally hosted model (*KPSCMI*), and zero-shot generation via clinical-situation abstraction and prompt optimization (*OptiMed*).

For Subtask 2 (Evidence Identification), most systems used both the patient-authored and clinician-interpreted questions as input (11/14). Zero-shot prompting was marginally more common than few-shot prompting (7/14 vs. 6/14), and API-based models dominated (10/14). Methodologically, most submissions framed evidence identification either as direct extraction of supporting sentence IDs from the note excerpt (10/14) or as sentence-wise classification (3/14). Verification-based strategies were common: five systems used self-reflection or iterative refinement, two used self-consistency across multiple generations, and four used multi-model ensembling. Three systems also reported prompt optimization, while one system departed from the dominant prompting-based methods by

using embedding-based sentence similarity. A recurring theme was recall optimization: several systems used inclusive prompts, answer-first reasoning, union voting, or post-hoc refinement to reduce the risk of omitting essential evidence. The highest-ranked systems illustrate two contrasting but effective strategies: calibrated sentence-level classification with verification and self-consistency (*Neural*), and zero-shot multi-model ensembling with majority voting (*OptiMed*).

For Subtask 3 (Answer Generation), both few-shot and zero-shot prompting were common, with few-shot setups only marginally more common than zero-shot setups (6/11 vs. 5/11). API-based models dominated this subtask (9/11), and most systems used both the patient-authored and clinician-interpreted questions at inference time (9/11). A common design choice was to condition generation on a filtered evidence subset, often obtained from the team's Subtask 2 pipeline, rather than on the full note excerpt alone (6/11 systems). Two systems used model ensembles, three used self-reflection or iterative self-revision, and three optimized prompts on the development set. Several teams adopted multi-stage generation procedures, including candidate generation followed by LLM-based selection, generate–critique–revise pipelines, and cite–then–rewrite strategies. As in the other subtasks, explicit task-specific fine-tuning was rare. The strongest reported systems again reflect different design choices: dynamic example retrieval with few-shot prompting (*WisPerMed*), intent-consistent in-context example selection (*TAMU-NLP-Lab*), and evidence-grounded multi-model generation with LLM-based selection (*BIT.UA-AAUBS*).

For Subtask 4 (Evidence Alignment), few-shot prompting was somewhat more common than zero-shot prompting (8/14 vs. 6/14). Nine of the fifteen systems used API-based models, while the remainder relied exclusively on locally hosted models. Input configurations were more varied than in Subtask 3: 7/14 systems used both question forms, 4 used only the patient question, and 3 used only the clinician-interpreted question. There was greater emphasis on precision–recall calibration through voting thresholds, self-consistency, and prompt instructions that explicitly favored either more conservative or more inclusive linking. Three systems used majority-vote ensembling across models, five used self-consistency, and only two explicitly reported prompt optimization. A small number of systems supplemented LLM predictions with embedding-based heuristics or fine-tuned reranking components. The highest-ranked systems show that strong performance can be achieved with either few-shot majority-vote ensembling (*BIT.UA-AAUBS*), zero-shot prompt-only alignment (*WisPerMed*), or ensemble prediction augmented with embedding-based recall heuristics (*Yale-DM-Lab*).

Taken together, the submitted system descriptions suggest that the benchmark decomposition influenced not only evaluation but also system design. Several teams explicitly propagated information across stages, e.g., by feeding predicted evidence into answer generation or by separating evidence identification from answer generation and alignment.

6. Conclusion

We presented an overview of the ArchEHR-QA 2026 Shared Task on grounded question answering from electronic health records. By decomposing the problem into question interpretation, evidence identification, answer generation, and evidence alignment, the benchmark enables separate evaluation of the main capabilities required for grounded response drafting to patient questions. The shared task attracted 198 submitted runs from 43 teams, with 17 teams contributing system descriptions for this overview. The highest-ranked systems differed across subtasks, and the largest gains over the organizer baseline were observed for evidence identification and evidence alignment. Across submitted systems, prompt-based LLM pipelines were dominant, task-specific fine-tuning was rare, and the strongest evidence-centric systems commonly combined retrieval, self-consistency, iterative refinement, and ensembling. We expect ArchEHR-QA 2026 to support future work on clinically faithful and verifiable QA systems for patient–clinician communication.

7. Limitations

This study has several limitations. First, the official test sets are modest in size, especially for Subtasks 1–3 (47 cases), so small numerical differences between closely ranked systems should not be over-interpreted. Second, the qualitative method analysis is limited to teams that submitted system descriptions; the remaining participating teams are reflected in the overall participation counts but not in the approach analysis. Third, evaluation relies primarily on automatic metrics. Although these metrics provide broad coverage of lexical and semantic similarity, they do not fully capture clinical usefulness, communication quality, or patient safety. This issue is especially relevant for Subtasks 1 and 3, where a single reference output may not capture the full range of clinically acceptable responses, and reference-based metrics may therefore under-reward valid paraphrases. Finally, the benchmark evaluates subtasks independently and uses curated note excerpts rather than the full

longitudinal EHR. Consequently, it does not fully capture end-to-end error propagation, open-ended retrieval over complete patient records, or generalization to other care settings such as outpatient longitudinal messaging.

8. Acknowledgements

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH) and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services. We thank all participating teams for their submissions and for contributing system descriptions used in this overview.

9. Bibliographical References

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2025. GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning. In *First Workshop on Foundations of Reasoning in Language Models*.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 Shared Task on Summarization in the Medical Domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85, Online. 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.

Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park. 2006. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 228–235, New York, NY, USA. Association for Computing Machinery.

Gregory Kell, Angus Roberts, Serge Umansky, Linglong Qian, Davide Ferrari, Frank Soboczenski, Byron C Wallace, Nikhil Patel, and Iain J Marshall. 2024. Question answering systems for health professionals at the point of care—a systematic review. *Journal of the American Medical Informatics Association*, 31(4):1009–1024.

Richelle J. Koopman, Linsey M. Barker Steege, Joi L. Moore, Martina A. Clarke, Shannon M. Canfield, Min S. Kim, and Jeffery L. Belden. 2015. Physician Information Needs and Electronic Health Records (EHRs): Time to Reengineer the Clinic Note. *The Journal of the American Board of Family Medicine*, 28(3):316–323.

Holly Jordan Lanham, Luci K. Leykum, and Jacqueline A. Pugh. 2018. Examining the Complexity of Patient-Outpatient Care Team Secure Message Communication: Qualitative Analysis. *Journal of Medical Internet Research*, 20(7):e9269.

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.

Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David R Karger. 2003. What Makes a Good Answer? The Role of Context in Question Answering. In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction*, Zurich, Switzerland.

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.

Sohn Nijor, Gavin Rallis, Nimit Lad, and Eric Gokcen. 2022. Patient Safety Issues From Information Overload in Electronic Medical Records. *Journal of Patient Safety*, 18(6):e999.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing Instructions and Demonstrations for Multi-Stage Language Model Programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366, Miami, Florida, USA. Association for Computational Linguistics.

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.

Sonish Sivarajkumar, Haneef Ahamed Mohammad, David Oniani, Kirk Roberts, William Hersh, Hongfang Liu, Daqing He, Shyam Visweswaran, and Yanshan Wang. 2024. Clinical Information Retrieval: A Literature Review. *Journal of Healthcare Informatics Research*, 8(2):313–352.

Sarvesh Soni and Dina Demner-Fushman. 2025a. Automated Evaluation can Distinguish the Good and Bad AI Responses to Patient Questions about Hospitalization.

Sarvesh Soni and Dina Demner-Fushman. 2025b. Automatically Generating Patient-specific Clinical Questions from Patient Messages to Relieve Clinician Burden. In *AMIA Annual Symposium Proceedings*, Atlanta, Georgia. American Medical Informatics Association.

Sarvesh Soni and Dina Demner-Fushman. 2026. A Dataset for Addressing Patient's Information Needs related to Clinical Course of Hospitalization. *Scientific Data*.

Sarvesh Soni, Soumya Gayen, and Dina Demner-Fushman. 2025. Overview of the ArchEHR-QA 2025 Shared Task on Grounded Question Answering from Electronic Health Records. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 396–405, Viena, Austria. Association for Computational Linguistics.

Ming Tai-Seale, Cliff W. Olson, Jinnan Li, Albert S. Chan, Criss Morikawa, Meg Durbin, Wei Wang, and Harold S. Luft. 2017. Electronic Health Record Logs Indicate That Physicians Split Time Evenly Between Seeing Patients And Desktop Medicine. *Health Affairs*, 36(4):655–662.

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.

Qi Yan, Zheng Jiang, Zachary Harbin, Preston H Tolbert, and Mark G Davies. 2021. Exploring the relationship between electronic health records and provider burnout: A systematic review. *Journal of the American Medical Informatics Association*, 28(5):1009–1021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report.

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.

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

```
You are an expert clinical natural language processing (NLP) assistant. Your task is to transform a
patient-authored question into a clear and concise clinician-interpreted question.

You are given:
- a patient-authored question.

Your goal:
- Write a clinician-interpreted question that captures the core clinical information need implied by the
patient-authored question.
- Phrase it as a query a clinician would ask a smart electronic health record (EHR) system to retrieve
relevant chart information needed to answer the patient.
- Keep it concise (maximum 15 words).
- Preserve the patient's core concern without introducing new clinical facts not explicitly stated in the
patient's narrative.

Output only the clinician-interpreted question, nothing else.

Patient question:
{___}
```

Table 8: Zero-shot prompt used for the Subtask 1 baseline.

```
You are an expert clinical natural language processing (NLP) assistant. Your task is to select the sen-
tence(s) in a clinical note excerpt that contain the clinical evidence needed to answer a patient's ques-
tion.

You are given:
- A patient-authored question.
- A clinician-interpreted version of the patient question.
- The clinical specialty(ies) relevant to the patient's question (may differ from the note's specialty;
comma-separated).
- A clinical note excerpt with numbered sentences.

Your goal:
- Select sentence IDs from the note excerpt that provide sufficient clinical evidence to answer the pa-
tient's question.
- Include all sentences necessary to fully support the answer (e.g., event + date; medication name +
dose).
- Only include sentences that are directly relevant and answer-bearing, including necessary qualifiers
(e.g., dates, values, negations, attribution).
- Do not include sentences that are unrelated, generic, purely administrative, or only tangentially re-
lated.
- If there is no sentence in the excerpt that provides evidence to answer the question, output an empty
array: [].
- You may select as many sentences as needed to provide sufficient evidence; the entire excerpt may not be
required.

Output only a JSON array of sentence ID numbers (as integers), nothing else. For example: [1, 5, 6]

Patient question:
{___}

Clinician-interpreted question:
{___}

Clinical specialty(ies) of the question (comma-separated):
{___}

Clinical note excerpt (with numbered sentences):
{___}
```

Table 9: Zero-shot prompt used for the Subtask 2 baseline.

```
You are an expert clinical natural language processing (NLP) assistant.  Your task is to generate a concise, clinically grounded answer to a patient's question using only the information in the provided clinical note excerpt.

You are given:
- A patient-authored question.
- A clinician-interpreted version of the patient question.
- The clinical specialty(ies) relevant to the patient's question (may differ from the note's specialty; comma-separated).
- A clinical note excerpt.

Your goal:
- Answer the patient's question using only information found in the clinical note excerpt.
- Write in a professional clinical register (not simplified lay language).
- Be concise:  limit your answer to 75 words (approximately 5 sentences).
- Stay faithful to the clinical note -- do not speculate or add information not supported by the note.
- If the note does not contain sufficient information to fully answer the question, provide a faithful response based on what is available without speculation.
- Do not use outside medical knowledge or inference.  Only use what is explicitly stated in the note excerpt.

Output only the answer text, nothing else.

Patient question:
{___}

Clinician-interpreted question:
{___}

Clinical specialty(ies) of the question (comma-separated):
{___}

Clinical note excerpt:
{___}
```

Table 10: Zero-shot prompt used for the Subtask 3 baseline.

```
You are an expert clinical natural language processing (NLP) assistant.  Your task is to align each answer sentence to the clinical note sentence(s) that support it.

You are given:
- A clinician-interpreted version of the patient question.
- A clinical note excerpt with numbered sentences.
- An answer with numbered sentences.

Your goal:
- For each answer sentence, select the clinical note sentence(s) that directly support it.
- If an answer sentence is supported by multiple note sentences, include all supporting note sentences. Assign an empty list only if that answer sentence is not supported by any note sentence.
- Each answer sentence may be supported by zero, one, or multiple note sentences.

Output a JSON array where each element has "answer_id" (string) and "evidence_id" (array of strings). Output nothing else.

Example output format:
[{"answer_id":  "1", "evidence_id":  ["2"]}, {"answer_id":  "2", "evidence_id":  ["5", "6"]}, {"answer_id":  "3", "evidence_id":  []}]

Clinician-interpreted question:
{___}

Clinical note excerpt (numbered sentences):
{___}

Answer (numbered sentences, no citations):
{___}
```

Table 11: Zero-shot prompt used for the Subtask 4 baseline.