# TAMU-NLP at ArchEHR-QA 2026: Grounded Clinical QA with Evidence Identification and Intent-Aware Answer Generation

**Xinqi Su**[1], **Rongrong Wang**[2], **Sunyang Fu**[2], **Hongfang Liu**[2], **Ruihong Huang**[1]
[1]Texas A&M University, College Station, TX, USA
[2]UTHealth Houston, Houston, TX, USA
{suxinqi666, huangrh}@tamu.edu
{Rongrong.Wang, Sunyang.Fu, Hongfang.Liu}@uth.tmc.edu

**Abstract**

Electronic Health Records (EHRs) contain rich clinical information and provide an important data source for medical question answering. However, generating reliable answers grounded in patient-specific clinical evidence remains challenging. In this work, we participate in the ArchEHR-QA 2026 shared task and focus on Subtask 2 (Evidence Identification) and Subtask 3 (Answer Generation). For evidence identification, we explore both traditional learning-to-rank methods and large language models (LLMs), and propose a two-stage LLM framework that improves prediction stability through few-shot prompting and self-reflection reasoning. For answer generation, we design an intent-aware few-shot prompting framework to generate concise answers grounded in clinical evidence. Experimental results show that our approach achieves strong performance despite limited training data. On the official leaderboard, our system ranks 5th in Subtask 2 and 2nd in Subtask 3. These results demonstrate that combining evidence-driven reasoning with the generative capabilities of LLMs is an effective approach for EHR-based clinical question answering.

**Keywords:** Electronic Health Records (EHR), Clinical Question Answering, Evidence Identification, Answer Generation

## 1. Introduction

Electronic Health Records (EHRs) digitally store and manage patient health information, including medical history, laboratory results, medications, and clinical procedures, providing essential data for clinical decision-making and research. With the widespread adoption of EHR systems and patient portals, clinicians receive increasing numbers of patient inquiries, many of which arise from patients reviewing and questioning their health information (Cronin et al., 2015; Irizarry et al., 2015). These inquiries not only increase clinicians' workload but also consume substantial clinical resources.

Although prior work has explored automated health question answering (QA), such as clinical and biomedical QA systems (Ben Abacha and Demner-Fushman, 2019; Pampari et al., 2018), relatively limited attention has been paid to leveraging patient-specific clinical records and grounding answers in explicit clinical evidence. Generating personalized, evidence-supported responses from EHR data therefore represents a promising direction for improving clinical efficiency and providing patients with reliable health information.

The ArchEHR-QA 2026 shared task evaluates EHR-based question answering systems through four subtasks emphasizing evidence grounding and clinical interpretability. Subtask 1 (Question Interpretation) converts patient free-text questions into concise clinical queries understandable to clinicians. Subtask 2 (Evidence Identification) extracts key evidence sentences from clinical notes required to answer the question. Subtask 3 (Answer Generation) produces concise answers grounded in the selected evidence. Subtask 4 (Evidence Alignment) evaluates the alignment between answers and their supporting evidence.

In this work, we focus on Subtasks 2 and 3, which form a natural reasoning pipeline in EHR question answering: evidence identification determines the information basis for answer generation, while answer generation produces patient-oriented explanations grounded in the selected evidence (Ben Abacha and Demner-Fushman, 2019).

For Subtask 2, we explore both traditional learning-to-rank methods (Cao et al., 2007; Burges, 2010) and large language models (LLMs) for evidence identification. We propose a two-stage LLM-based evidence identification framework, which performs initial predictions using few-shot prompting and improves prediction stability through self-reflection reasoning and ensemble voting.

For Subtask 3, we adopt a few-shot intent-aware prompting framework based on LLMs to generate concise and evidence-grounded answers. By leveraging the reasoning capability of LLMs, the system can produce answers consistent with clinical evidence even under limited training data.

Our approach achieves competitive performance. On the official ArchEHR-QA 2026 leaderboard, our system ranks 5th in Subtask 2 (Evidence Identification) and 2nd in Subtask 3 (Answer Generation). These results demonstrate that combining evidence-driven reasoning strategies with the generative capabilities of large language models provides an effective solution for EHR-based clinical question answering.



*Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 551–564
12 May 2026. 

## 2. Related Work

### 2.1. Evidence Identification

Evidence identification aims to extract a minimal set of sentences from clinical notes that support answer generation. Early approaches relied on rule-based systems or keyword matching, such as evidence extraction using regular expressions or ontology-based medical terms (Demner-Fushman and Lin, 2007). With the development of deep learning, supervised sentence-level models were introduced, treating evidence identification as a binary classification task to determine whether a candidate sentence supports the answer (Chakraborty et al., 2020). These models often extract entities and relations as features by performing information extraction tasks including clinical named entity recognition (NER) and relation extraction (RE). Recently, large language models (LLMs) have been applied to evidence identification, enabling zero-shot or few-shot evidence prediction through strong semantic understanding (Singhal et al., 2023). However, most existing studies focus on biomedical literature or general medical QA, while evidence retrieval from patient-specific EHR data remains underexplored, particularly under small datasets where balancing precision and recall is challenging.

### 2.2. Answer Generation

The goal of answer generation is to produce concise and accurate responses grounded in evidence. Early medical QA systems mainly relied on information retrieval or template-based generation, where retrieved evidence snippets were directly assembled into answers, often lacking fluency and flexibility (Ben Abacha and Zweigenbaum, 2015; Ben Abacha and Demner-Fushman, 2019). With the development of deep learning, neural generation models have been increasingly applied, learning relationships among questions, evidence, and answers to produce more natural responses (Pampari et al., 2018; Jin et al., 2019; Lee et al., 2020). More recently, large language models (LLMs) have demonstrated strong capabilities, generating coherent responses under zero-shot or few-shot settings (Singhal et al., 2023; Nori et al., 2023). However, most existing studies focus on medical exam questions or biomedical literature, while answer generation for patient-specific EHR data remains underexplored, particularly under small-scale training settings where maintaining both answer reliability and evidence traceability is challenging.

## 3. Data and Evaluation Metrics

### 3.1. Data

The ArchEHR-QA 2026 dataset (Soni and Demner-Fushman, 2026b,a) is organized at the case level. The development set contains 20 cases (Cases 1–20). Each case consists of the following components:

- **Patient Narrative**: a free-text description of the patient's symptoms or health concerns from the patient perspective.
- **Clinician Question**: a concise clinical question formulated by a clinician based on the patient narrative.
- **Patient–Question Phrases**: key phrases in the narrative annotated as relevant to the patient's question.
- **Clinical Note Excerpts**: sentence-level excerpts from clinical notes, where each sentence is labeled as *essential*, *supplementary*, or *not relevant*. The notes are derived from the MIMIC database.
- **Clinician Answer**: a reference answer written by a clinician. Each answer sentence is supported by zero or more sentences from the clinical note excerpts.
- **Case Metadata**: metadata including the case ID, clinical specialty, and EHR source.

In addition to the development set, the dataset includes Cases 21–120, which lack annotated evidence labels for the clinical note excerpts. To address the lack of annotated training data, we employ a large language model (LLM) to automatically annotate evidence sentences from the clinical note excerpts based on the clinician answers. The resulting LLM-annotated data are used as training data for Task 2 (Evidence Identification), while the original development set (Cases 1–20) is used as the validation set. Furthermore, the official test set for Task 2 and Task 3 includes 47 additional cases (Cases 121–167), which follow a structure similar to the development set but do not include annotated evidence labels or clinician answers, and are used solely for evaluation.

Although the original annotations include three relevance labels, we cast the task as a binary classification problem since Task 2 evaluation and Task 3 clinician answers primarily rely on *essential* evidence. Specifically, only *essential* sentences are treated as gold evidence, while the other two categories are merged into *not-essential*. In the development set, the label distribution consists of 121 *essential* and 307 *not-essential* sentences, reflecting a class imbalance.

### 3.2. Evaluation Metrics

**Task 3: Answer Generation.** We evaluate answer generation using a diverse set of automatic

metrics that capture different aspects of answer quality. **BLEU** and **ROUGE** measure n-gram overlap between generated and reference answers. **BERTScore** evaluates semantic similarity between generated and reference texts. **SARI** (Xu et al., 2016) measures text editing quality by comparing added, deleted, and retained n-grams with respect to the reference. **AlignScore** (Rogers et al., 2023) assesses semantic and factual alignment using a trained scoring model that evaluates the consistency between generated and reference answers. **MEDCON** (Yim et al., 2023) evaluates medical consistency by checking whether clinically relevant information is preserved without contradiction. Together, these metrics provide a comprehensive evaluation of lexical similarity, semantic alignment, content faithfulness, and clinical correctness.

## 4. Methods

In this section, we describe in detail the approaches we explored for Task 2 and Task 3.

### 4.1. Task 2: Evidence Identification

#### 4.1.1. Task Formulation

Given a clinician question $q$, a patient narrative $p$, and a set of candidate sentences extracted from clinical notes $S = \{s_1, s_2, \ldots, s_n\}$, the goal of evidence identification is to identify sentences that provide key evidence for answering the question.

Although the original annotations contain three labels (*essential*, *supplementary*, and *not-essential*), only *essential* sentences are treated as gold evidence during evaluation. We therefore formulate the task as a sentence-level relevance prediction problem: $y_i \in \{0, 1\}$, where $y_i = 1$ indicates that sentence $s_i$ is an essential evidence sentence.

Since each case contains multiple candidate sentences, we further model the task as a sentence ranking problem by learning a scoring function

$$\text{score}_i = f(q, p, s_i), \quad s_i \in S \tag{1}$$

such that evidence sentences receive higher scores than non-evidence sentences.

#### 4.1.2. Ranking-based Evidence Identification

**Evidence Ranking Model** To model the relevance between the question and candidate sentences, we adopt a cross-encoder architecture. Unlike bi-encoders that encode the query and sentence independently, the cross-encoder jointly encodes the question–sentence pair, enabling fine-grained token-level interactions (Reimers and Gurevych, 2019; Nogueira and Cho, 2019). Since each case contains only a small number of candidate sentences, the additional computational cost remains manageable.

We initially experimented with incorporating the clinical specialty and the patient narrative $p$ into the query representation. However, empirical results show little improvement compared with using only the clinician question $q$. This is likely because the clinician question already summarizes the key information from the patient narrative, while the narrative often contains additional background details that are less relevant for evidence selection. Therefore, we use only the question $q$ as the query input in the final model. For each question–sentence pair $(q, s_i)$, the input sequence is constructed as $x_i = [\text{CLS}]\ q\ [\text{SEP}]\ s_i\ [\text{SEP}]$. The sequence is fed into an encoder $f_\theta$, and the hidden representation of the [CLS] token is used as the sentence-pair representation:

$$h_i = f_\theta(x_i)_{[\text{CLS}]} \tag{2}$$

A linear layer is then applied to compute the relevance score:

$$\text{score}_i = W h_i + b \tag{3}$$

**Learning Objectives** We explore three learning strategies: **pointwise classification**, **pairwise ranking**, and **listwise ranking**, and further investigate **hybrid objectives** that combine classification and ranking signals.

**Pointwise Classification** We first model the task as a sentence-level binary classification problem. The probability of a sentence being evidence is computed as

$$p_i = \sigma(\text{score}_i). \tag{4}$$

The training objective is the binary cross-entropy (BCE) loss:

$$L_{\text{point}} = -\sum_i \left[ y_i \log p_i + (1 - y_i) \log(1 - p_i) \right]. \tag{5}$$

Since evidence sentences are much fewer than non-evidence sentences, we apply weighted BCE loss to alleviate class imbalance (Lin et al., 2017).

**Pairwise Ranking** To learn the relative ordering of sentences, we adopt pairwise learning-to-rank. For each case, we construct positive–negative sentence pairs $(s^+, s^-)$. To reduce training difficulty, we perform hard negative mining, selecting the top-$k$ hardest negative sentences (i.e., those with the highest predicted scores). The margin ranking loss is defined as:

$$L_{\text{pair}} = \max(0, m - \text{score}_{pos} + \text{score}_{neg}). \tag{6}$$

This objective encourages

$$\text{score}_{pos} > \text{score}_{neg} + m. \tag{7}$$

thereby improving the model's ability to distinguish evidence sentences from hard negatives.

**Listwise Ranking** We also explore listwise ranking, which learns sentence ranking at the case level. Given model scores $\text{score}_i$, we compute a probability distribution:

$$P_i = \frac{\exp(\text{score}_i)}{\sum_j \exp(\text{score}_j)}. \tag{8}$$

The target distribution assigns probability mass only to gold evidence sentences:

$$G_i = \frac{y_i}{\sum_j y_j}. \tag{9}$$

The listwise loss is defined as

$$L_{\text{list}} = -\sum_i G_i \log P_i. \tag{10}$$

**Hybrid Objectives** To combine the benefits of classification and ranking signals, we further explore hybrid objectives:
**Pointwise + Pairwise:**

$$L = \alpha L_{\text{point}} + (1-\alpha) L_{\text{pair}}. \tag{11}$$

**Pointwise + Listwise:**

$$L = \alpha L_{\text{point}} + (1-\alpha) L_{\text{list}}. \tag{12}$$

where $\alpha$ is a weighting coefficient.

**Inference** During inference, the model computes a score $\text{score}_i$ or probability $p_i$ for each candidate sentence and applies several strategies to determine the number of evidence sentences.

**Fixed Threshold** We first select evidence sentences using a fixed probability threshold:

$$\hat{E} = \{s_i \mid p_i \geq \tau\}. \tag{13}$$

where the threshold $\tau$ is tuned on the development set to maximize Strict Micro-F1. If no sentence exceeds the threshold, the sentence with the highest score is selected.

**Score-Gap Selection** To adaptively determine the number of evidence sentences, we adopt a score-gap strategy. Candidate sentences are first sorted by their scores: $\text{score}_{(1)} \geq \text{score}_{(2)} \geq \cdots \geq \text{score}_{(n)}$. We then compute the score difference between adjacent sentences:

$$\text{gap}_i = \text{score}_{(i)} - \text{score}_{(i+1)}. \tag{14}$$

The largest gap typically indicates the boundary between evidence and non-evidence sentences:

$$k = \arg\max_i \ \text{gap}_i. \tag{15}$$

The top-$k$ sentences are selected as predicted evidence: $\hat{E} = \{s_{(1)}, \ldots, s_{(k)}\}$.

**Adaptive Soft Dynamic-$k$** We further propose an Adaptive Soft Dynamic-$k$ method that determines the number of evidence sentences based on prediction uncertainty. Sentence scores are first normalized with softmax:

$$p_i = \frac{\exp(\text{score}_i)}{\sum_j \exp(\text{score}_j)}. \tag{16}$$

We then compute the normalized entropy of the distribution:

$$H_{\text{norm}} = -\frac{\sum_i p_i \log p_i}{\log(n)}. \tag{17}$$

A dynamic threshold is defined as

$$\tau = \tau_0 + \lambda H_{\text{norm}}, \tag{18}$$

where $\tau_0$ and $\lambda$ are tuned on the development set. Sentences are sorted by probability and accumulated until

$$\sum_{i=1}^{k} p_i \geq \tau. \tag{19}$$

This yields a dynamic number of evidence sentences $k$. When the prediction distribution is concentrated (low entropy), only a few high-confidence sentences are selected; when it is more dispersed (high entropy), more sentences are retained. This helps preserve important evidence and provides more reliable support for answer generation.

#### 4.1.3. LLM-based Evidence Identification

To explore the capabilities of large language models (LLMs) for evidence identification, we design a two-stage LLM framework that combines retrieval-based few-shot prompting and self-reflection reasoning to improve prediction accuracy and stability.

**Retrieval-based Few-shot Prompting** To guide the model toward consistent annotation patterns, we adopt retrieval-based few-shot prompting. Instead of randomly selecting examples, we retrieve several annotated cases from the development set based on semantic similarity to the current query.

Each example contains the patient narrative, clinician question, clinical note, and labels. These examples provide explicit supervision signals that help the model determine whether a sentence constitutes key evidence.

**Stage 1: Initial Evidence Prediction** In the first stage, we use Gemini-2.5-Pro to perform initial evidence identification. The model receives the retrieved few-shot examples, patient narrative, clinician question, and clinical note, and predicts structured labels for each sentence:

$$Y^{(1)} = \{y_1^{(1)}, y_2^{(1)}, ..., y_n^{(1)}\}. \tag{20}$$

Each sentence is labeled as *essential* or *not-essential*.

**Stage 2: Self-Reflection Refinement** Single-pass LLM predictions may miss important evidence or over-predict irrelevant sentences. To mitigate this issue, we introduce a self-reflection stage. The model receives the initial prediction $Y^{(1)}$ and re-evaluates its decisions by checking whether key medical evidence was missed or incorrectly labeled. After reflection, the model produces refined predictions:

$$Y^{(2)} = \{y_1^{(2)}, y_2^{(2)}, ..., y_n^{(2)}\}. \tag{21}$$

The final output is defined as $Y = Y^{(2)}$.

**Ensemble Voting** To further improve robustness, we perform ensemble voting over multiple LLM predictions. For each sentence, we count how many times it is labeled as *essential*. A sentence is predicted as essential if its vote count exceeds a predefined threshold $T$:

$$y_i = \begin{cases} 1 & v_i \geq T \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

where $v_i$ denotes the number of *essential* votes. Adjusting $T$ allows the system to flexibly balance precision and recall in evidence identification.

## 4.2. Task 3: Answer Generation

Given a clinician question $q$, a patient narrative $n$, and the evidence $E = \{e_1, e_2, \ldots, e_k\}$ identified in Section 4.1, the goal of answer generation is to produce an answer $a$ grounded in the evidence.

Clinical questions often reflect different information needs, such as treatment planning, prognosis assessment, or interpretation of abnormal laboratory results. To better support clinical reasoning and improve answer quality, we incorporate question intent information and retrieve intent-aligned demonstrations to guide generation. This intent-aware strategy provides reasoning patterns that match the question type, improving consistency between generated answers and clinical evidence. Our framework consists of three stages: (1) intent identification, (2) intent-aware example retrieval, and (3) evidence-grounded answer generation.

### 4.2.1. Intent Identification

We formulate intent prediction as a classification task. For each case $x = (n, q)$, we define a taxonomy of 12 clinical question intents (e.g., treatment plan, medication safety, prognosis, and laboratory abnormality explanation). Given a question, the model predicts a primary intent $I$ and optionally a secondary intent $S$. The predicted intents are generated by the DeepSeek-V3 model, yielding a structured representation $x = (n, q, I, S)$.

### 4.2.2. Intent-Aware Example Retrieval

To support in-context learning, we construct an example pool $P$ with intent annotations. For a test sample $x_i$ with primary intent $I_i$ and secondary intent $S_i$, we first retrieve candidate examples that share the same primary intent:

$$C_i = \{x_j \in P \mid I_j = I_i\}. \tag{23}$$

If the number of retrieved candidates is fewer than $K$, we further include examples that match the secondary intent $S_i$. The example pool is constructed from Cases 1–120, where each case is annotated with intent labels. The pool covers all predefined intent categories, ensuring that each intent type has sufficient example support.

We then rank candidates using semantic similarity. Each sample is represented as $t = I \oplus q \oplus n$, where $\oplus$ denotes text concatenation. Let $\mathbf{v}_q$ and $\mathbf{v}_j$ denote the embeddings of the query and candidate example. Similarity is computed as

$$\text{sim}(q, x_j) = \frac{\mathbf{v}_q \cdot \mathbf{v}_j}{\|\mathbf{v}_q\|\|\mathbf{v}_j\|}. \tag{24}$$

The top-$K$ most similar examples are selected:

$$D_i = \text{TopK}_{x_j \in C_i} \text{sim}(x_i, x_j). \tag{25}$$

### 4.2.3. Evidence-Grounded Answer Generation

Finally, we generate answers using Gemini-2.5-Pro conditioned on the question, evidence sentences, intent information, and retrieved demonstrations. The generation function can be written as

$$a = g(n, q, E, D). \tag{26}$$

Each demonstration in $D$ contains the patient narrative, clinician question, evidence sentences, and a reference answer, illustrating clinically grounded reasoning. Incorporating evidence and intent information encourages the model to generate answers consistent with the supporting evidence.

# 5. Results

In this section, we present the implementation details of our approaches and report experimental results on the development set as well as the official evaluation results.

## 5.1. Implementation Details

All experiments are implemented in PyTorch with the Transformers library and run on two NVIDIA RTX 3090 GPUs. For ranking-based evidence

| Method | Strict | | | | | | Lenient | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Macro | | | Micro | | | Macro | | | Micro | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Pointwise | 0.5113 | 0.3692 | 0.3886 | 0.5000 | 0.3471 | 0.4098 | 0.5472 | 0.3364 | 0.3826 | 0.5476 | 0.3129 | 0.3983 |
| Pairwise Ranking | 0.4463 | **0.8566** | 0.5594 | 0.3820 | **0.8430** | 0.5258 | 0.5201 | **0.8225** | **0.6104** | 0.4532 | **0.8231** | 0.5845 |
| Listwise Ranking | 0.4750 | 0.7205 | 0.5462 | 0.4750 | 0.6281 | 0.5409 | 0.5437 | 0.6920 | 0.5820 | 0.5437 | 0.5918 | 0.5668 |
| Pointwise + Pairwise | 0.5214 | 0.6758 | **0.5617** | 0.5214 | 0.6033 | **0.5594** | 0.6000 | 0.6634 | 0.6017 | 0.6000 | 0.5714 | **0.5854** |
| Pointwise + Listwise | **0.5500** | 0.6278 | 0.5554 | **0.5500** | 0.5455 | 0.5477 | **0.6250** | 0.5881 | 0.5781 | **0.6250** | 0.5102 | 0.5618 |

Table 1: Results on Task 2: Evidence Identification on the development set under Strict and Lenient evaluation settings. We report Macro and Micro Precision (P), Recall (R), and F1.

| Method | P | R | F1 |
|---|---|---|---|
| Fixed Threshold | 0.5214 | 0.6033 | 0.5594 |
| Score-Gap Selection | 0.4875 | **0.6446** | 0.5552 |
| Adaptive Soft Dynamic-$k$ | **0.5469** | 0.5800 | **0.5623** |

Table 2: Effect of inference strategies on Task 2 under the Strict setting (development set). We report Precision (P), Recall (R), and Micro-F1.

identification, we adopt a cross-encoder architecture based on `BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext`. We also experiment with several alternative backbones, including `DeBERTa-v3-large`, `DeBERTa-v3-base`, `BioClinicalBERT`, and `BioBERT-base-cased-v1.2`. However, these models yield similar performance on the development set, so we report the results using PubMedBERT.

The model is trained for 10 epochs using the AdamW optimizer with a learning rate of $2 \times 10^{-5}$ and weight decay of $0.01$. The maximum input length is set to 384, and gradient clipping with a norm of 1.0 is applied. For the hybrid loss, the weight $\alpha$ is set to 0.1 and margin is set to 0.2. To ensure reproducibility, all experiments use a fixed random seed of 42. For LLM-based methods, we access the models through the official API.

## 5.2. Task 2: Evidence Identification

### 5.2.1. Effect of Ranking Objectives on Evidence Identification

Table 4.2.3 reports the performance of different evidence identification methods on the development set under both Strict and Lenient evaluation settings, including Macro and Micro Precision, Recall, and F1. All experiments use a fixed threshold (0.4) during inference for evidence selection. The results show that pointwise classification performs the worst overall. Although it achieves relatively high Micro Precision (0.5000) under the Strict setting, its Recall is much lower (0.3471), resulting in a Strict Micro-F1 of only 0.4098. This suggests that modeling evidence identification as independent sentence-level classification is insufficient to recover all key evidence sentences, since the task inherently requires comparing multiple candidate sentences within the same case.

In contrast, ranking-based methods substantially improve recall. Pairwise ranking achieves the highest Strict Micro Recall (0.8430), while listwise ranking also outperforms pointwise classification. This indicates that modeling the relative relevance among candidate sentences within each case better matches the nature of the task.

Combining classification and ranking objectives further improves the precision–recall balance. For example, Pointwise + Pairwise increases Precision while maintaining high Recall, and Pointwise + Listwise achieves the best Strict Micro-F1 (0.5477). Similar trends are observed under the Lenient setting, where Pointwise + Pairwise obtains the highest Lenient Micro-F1 (0.5854).

Overall, these results suggest that modeling evidence identification as a ranking problem is more suitable, and integrating classification and ranking signals further improves performance, particularly under limited training data.

### 5.2.2. Inference Strategy Analysis

Table 4.2.3 compares three evidence selection strategies under the Strict evaluation setting: Fixed Threshold, Score-Gap Selection, and Adaptive Soft Dynamic-$k$. All strategies are applied on top of the Pointwise + Pairwise model during inference. Although the overall F1 scores are similar, the methods exhibit different precision–recall trade-offs.

The Fixed Threshold method achieves relatively higher recall but lower precision, indicating that a fixed threshold tends to retain more candidate sentences, increasing recall while introducing more false positives. Score-Gap Selection achieves the highest recall (0.6446) but the lowest precision (0.4875). This is consistent with the intuition behind the method: automatically determining the number of evidence sentences based on score gaps tends to retain more potentially relevant sentences, which improves recall but also increases false positives.

In contrast, Adaptive Soft Dynamic-$k$ achieves the best Strict Micro-F1 and the highest precision while maintaining stable recall. This suggests that dynamically adjusting the number of evidence

| Method | Strict | | | | | | Lenient | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Macro | | | Micro | | | Macro | | | Micro | | |
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Few-shot | 0.8026 | 0.8269 | 0.7917 | 0.7692 | 0.8264 | 0.7968 | 0.8231 | 0.7060 | 0.7351 | 0.8000 | 0.7075 | 0.7509 |
| Few-shot + Self-reflection | **0.8523** | **0.8427** | **0.8252** | 0.7923 | **0.8512** | **0.8207** | **0.8773** | **0.7229** | **0.7679** | **0.8385** | **0.7415** | **0.7870** |
| Few-shot + Self-reflection + Voting | 0.8492 | 0.8416 | 0.8229 | **0.7969** | 0.8430 | 0.8193 | 0.8702 | 0.7169 | 0.7610 | 0.8359 | 0.7279 | 0.7782 |

Table 3: Ablation study of the LLM-based evidence identification framework on Task 2 evaluated on the development set. We report Macro and Micro Precision (P), Recall (R), and F1.

| Method | P | R | F1 |
|---|---|---|---|
| Qwen3-Max | 0.6944 | 0.8264 | 0.7547 |
| Deepseek-V3 | 0.6809 | 0.7934 | 0.7328 |
| Deepseek-R1 | 0.7500 | 0.7934 | 0.7711 |
| GPT-4.1 | 0.7966 | 0.7769 | 0.7866 |
| GPT-5.1 | **0.8103** | 0.7769 | 0.7932 |
| Gemini-2.5-Pro | 0.7923 | **0.8512** | **0.8207** |

Table 4: Effect of different LLMs on Task 2 on the development set under the Strict evaluation setting. We report Precision (P), Recall (R), and Micro F1.

sentences based on prediction uncertainty better adapts to varying evidence distributions across cases, resulting in a more balanced performance.

Overall, these results indicate that the evidence selection strategy during inference has a substantial impact on performance, and uncertainty-based dynamic selection provides a better balance between precision and recall.

#### 5.2.3. LLM-based Evidence Identification

Table 4.2.3 presents results of the LLM-based evidence identification, including Few-shot prompting (with $K = 5$ retrieved in-context examples), Few-shot + Self-reflection, and Few-shot + Self-reflection + Voting. Few-shot prompting alone achieves strong performance. Under the Strict setting, it obtains a Micro-F1 of 0.7968, substantially outperforming traditional ranking models. This suggests that structured in-context examples effectively guide LLMs to identify key evidence sentences.

Adding self-reflection further improves performance. Under the Strict setting, Micro-F1 increases from 0.7968 to 0.8207, with gains in both precision and recall. Similar improvements are observed under the Lenient setting. This indicates that the reflection stage allows the model to revisit its initial predictions and correct missing or redundant evidence. In contrast, adding voting after self-reflection does not yield further improvements. Although the results remain competitive, they are slightly lower than those obtained with self-reflection alone under both settings. A possible explanation is that predictions across runs tend to be highly similar, and majority voting may suppress some infrequent but important evidence sentences, slightly reducing recall.

Overall, these results indicate that the main performance gains of LLM-based evidence identification come from self-reflection reasoning, rather than simple multi-run voting.

#### 5.2.4. Comparison of LLM Backbones

Table 4.2.3 compares the performance of different LLM backbones on the evidence identification task. Gemini 2.5 Pro achieves the best results (Strict Micro-F1 = 0.8207) and the highest recall (0.8512), indicating stronger coverage in identifying clinically relevant evidence sentences. Given the imbalanced label distribution in the development set (121 *essential* vs. 307 *not-essential*), recall becomes particularly important for capturing sparse gold evidence. The higher recall of Gemini suggests that it is more effective at retrieving relevant evidence under this setting. The GPT models (GPT-4.1 and GPT-5.1) also perform well, with higher precision but slightly lower recall than Gemini, suggesting a more conservative evidence selection strategy that prioritizes precision over coverage. Among open-source models, DeepSeek-R1 performs best (Micro-F1 = 0.7711), while Qwen3-Max and DeepSeek-V3 perform slightly worse. This indicates that reasoning and instruction-following capabilities are particularly important for sentence-level evidence identification. Overall, stronger language understanding and reasoning abilities help models better distinguish key evidence from background information, especially under imbalanced label distributions.

### 5.3. Test Set Results

Since Few-shot + Self-reflection achieved the best performance on the development set, we adopt it as the final method for Subtask 2. Table 4.2.3 reports the official test results, where our system obtains an overall score of 61.4, ranking among the top systems in the shared task. Compared with higher-ranked systems such as Neural and OptiMed, our method maintains relatively high Strict Micro Recall (70.0) but slightly lower precision. This indicates

| Method | Overall | Strict Micro | | | Lenient Micro | | | Strict Macro | | | Lenient Macro | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| Neural | **63.7** | **60.2** | 67.6 | **63.7** | **77.8** | 67.6 | **72.4** | **64.3** | 69.7 | **64.8** | **81.9** | 69.7 | **73.1** |
| OptiMed | 63.2 | 56.7 | 71.3 | 63.2 | 73.1 | 71.3 | 72.2 | 61.5 | 73.0 | 64.1 | 78.1 | 73.0 | 72.9 |
| UIC-AIHealth4All | 62.9 | 59.3 | 67.0 | 62.9 | 74.3 | 67.0 | 70.4 | 61.8 | 69.2 | 63.0 | 77.8 | 69.2 | 70.5 |
| Yale-DM-Lab | 61.9 | 52.3 | **75.7** | 61.9 | 68.0 | **75.7** | 71.6 | 56.7 | **75.5** | 61.1 | 73.8 | **75.5** | 70.4 |
| TAMU-NLP-Lab (Ours) | 61.4 | 54.6 | 70.0 | 61.4 | 74.4 | 70.0 | 72.2 | 61.4 | 70.3 | 60.6 | 78.9 | 70.3 | 70.7 |

Table 5: Official leaderboard Results on Task 2 on the test set. We report Strict and Lenient Micro and Macro Precision (P), Recall (R), and F1 scores.

| Method | Overall | BLEU | ROUGE-Lsum | SARI | BERTScore | AlignScore | MEDCON |
|---|---|---|---|---|---|---|---|
| LLM + Evidence + Few-shot | 34.80 | 8.97 | 27.35 | 58.87 | 44.80 | 25.92 | **42.89** |
| LLM + Evidence + Few-shot Intent | **36.20** | **9.70** | **27.40** | **59.60** | **46.20** | **34.50** | 39.90 |

Table 6: Results on Task 3: Answer Generation on the test set. We report Overall score, BLEU, ROUGE-Lsum, SARI, BERTScore, AlignScore, and MEDCON.

| Method | Overall | BLEU | ROUGE-Lsum | SARI | BERTScore | AlignScore | MEDCON |
|---|---|---|---|---|---|---|---|
| LLM + Evidence + Intent (Low-Recall) | 35.74 | 8.54 | **27.81** | 59.48 | **46.77** | 31.85 | **39.98** |
| LLM + Evidence + Intent (High-Recall) | **36.20** | **9.70** | 27.40 | **59.60** | 46.20 | **34.50** | 39.90 |

Table 7: Results on Task 3: Answer Generation on the test set. We report Overall score, BLEU, ROUGE-Lsum, SARI, BERTScore, AlignScore, and MEDCON.

| Team | Overall | BLEU | ROUGELsum | SARI | BERTScore | AlignScore | MEDCON |
|---|---|---|---|---|---|---|---|
| WisPerMed | **36.3** | **9.9** | **27.8** | 58.6 | **46.8** | 31.7 | 43.1 |
| TAMU-NLP-Lab (Ours) | 36.2 | 9.7 | 27.4 | 59.6 | 46.2 | **34.5** | 39.9 |
| BIT.UA-AAUBS | 35.6 | 8.6 | 26.4 | **60.0** | 45.0 | 30.2 | **43.2** |
| Neural | 35.2 | 9.4 | 25.6 | 57.7 | 43.5 | 34.3 | 40.9 |
| HealthNLP_Retrievers | 34.6 | 7.0 | 25.4 | 59.2 | 43.8 | 33.6 | 38.7 |

Table 8: Official leaderboard results on ArchEHR-QA 2026 task 3 (Test Set).

that the system effectively covers key evidence sentences, although it may occasionally include some non-evidence sentences. Under the Lenient setting, our system achieves a Micro-F1 of 72.2 and a Lenient Macro-F1 of 70.7, suggesting stable performance across cases. Overall, the results show that our approach generalizes well to the shared task test set. Further improvements may come from enhancing precision in evidence selection.

## 5.4. Task 3: Answer Generation

### 5.4.1. Impact of Intent-Aware Few-Shot Retrieval

Table 4.2.3 compares two evidence-grounded answer generation settings: LLM + Evidence + Few-shot (with $K = 3$ retrieved in-context examples) and LLM + Evidence + Few-shot Intent. Both methods use evidence sentences and few-shot demonstrations, while the latter additionally retrieves demonstrations based on question intent. Introducing intent information improves the Overall score from 34.80 to 36.20, indicating that the selection of few-shot demonstrations plays an important role in clinical answer generation. Compared with generic examples, intent-matched demonstrations provide more effective guidance for generating high-quality responses. Across metrics, BLEU (8.97 → 9.70) and ROUGE-Lsum (27.35 → 27.40) show slight improvements, while BERTScore (44.80 → 46.20) and AlignScore (25.92 → 34.50) increase more substantially, suggesting better semantic alignment between generated and reference answers. In addition, SARI improves from 58.87 to 59.60, indicating better evidence integration and text restructuring. However, MEDCON slightly decreases (42.89 → 39.90), suggesting that improved semantic relevance may come at the cost of stricter medical terminology consistency. Overall, intent-aware retrieval of few-shot demonstrations significantly improves semantic alignment and overall answer quality in clinical answer generation.

### 5.4.2. Impact of Evidence Recall on Answer Generation

Table 4.2.3 compares the Task 3 answer generation results under two evidence input settings: Low-Recall Setting and High-Recall Setting. These two settings are constructed by controlling the precision–recall trade-off in the evidence identification stage. Specifically, the Low-Recall setting applies

a stricter evidence selection criterion to retain only high-confidence sentences, leading to higher precision but lower recall. In contrast, the High-Recall setting adopts a more permissive selection strategy, including more candidate sentences, which increases recall but reduces precision.

Both methods use the same LLM + Evidence + Intent framework and differ only in the coverage of the input evidence. Overall, the High-Recall Setting achieves a higher Overall score (36.20 vs. 35.74), indicating that broader evidence coverage generally improves the quality of generated clinical answers. Across individual metrics, this setting performs better on BLEU, SARI, and AlignScore, with the largest improvement observed in AlignScore (31.85 → 34.50). This suggests that, for LLMs, incorporating more evidence can significantly enhance the semantic alignment between generated answers and reference answers. In contrast, the Low-Recall Setting performs slightly better on ROUGE-Lsum, BERTScore, and MEDCON, suggesting that a more selective evidence set may sometimes produce answers that are more focused and contain more compact medical concepts.

#### 5.4.3. Leaderboard Results

Table 4.2.3 presents the official leaderboard results on the ArchEHR-QA 2026 Subtask 3 test set. Our system (TAMU-NLP-Lab) ranks second with an Overall score of 36.2, only slightly below WisPerMed (36.3), demonstrating strong competitiveness. Notably, our system achieves the highest AlignScore (34.5) among all teams, indicating that the generated answers are most semantically aligned with the reference answers. This suggests that combining intent-aware generation with evidence grounding helps the model better capture the core information needs of clinical questions. Our system also maintains competitive performance on BLEU (9.7) and ROUGE-Lsum (27.4), showing strong surface-level similarity with the reference answers. However, our system does not achieve the best results on SARI and MEDCON. For example, BIT.UA-AAUBS obtains the highest SARI (60.0), and both BIT.UA-AAUBS and WisPerMed slightly outperform our system on MEDCON. This suggests that while our approach focuses on addressing question intent, there is still room for improvement in text editing quality and strict medical concept consistency.

## 6. Conclusion

This paper presents an evidence-grounded framework for clinical question answering and conducts a comprehensive evaluation on the ArchEHR-QA dataset. The task is decomposed into two stages: evidence identification (Task 2) and answer generation (Task 3). For Task 2, we explore both ranking-based models and LLM-based evidence annotation. For Task 3, we introduce an intent-aware answer generation framework that retrieves few-shot demonstrations based on question intent to guide evidence-consistent responses. Experimental results show that incorporating LLM-based self-reflection significantly improves evidence identification, especially under limited training data, outperforming traditional ranking models. For answer generation, we find that evidence coverage and intent-aware few-shot retrieval are critical: higher evidence recall improves answer quality, while intent-based example retrieval enhances semantic alignment between generated and reference answers. In the ArchEHR-QA 2026 Shared Task, our system ranks 5th in Subtask 2 and 2nd in Subtask 3, achieving competitive performance across multiple metrics. Future work will explore tighter integration between evidence identification and answer generation to further improve end-to-end clinical QA performance.

## 7. Bibliographical References


Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511.

Asma Ben Abacha and Pierre Zweigenbaum. 2015. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information Processing & Management*, 51(5):570–594.

Christopher Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11:81.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of ICML*, pages 129–136.

Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. 2020. Biomedbert: A pre-trained biomedical language model for qa and ir. In *Proceedings of COLING*, pages 669–679.

Robert M Cronin, SE Davis, JA Shenson, Q Chen, ST Rosenbloom, and GP Jackson. 2015. Growth of secure messaging through a patient portal as a form of outpatient interaction across clinical specialties. *Applied Clinical Informatics*, 6(02):288–304.

Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103.

Taya Irizarry, Annette DeVito Dabbs, and Christine Curran. 2015. Patient portals and patient engagement: a state of the science review. *Journal of Medical Internet Research*, 17(6):e148.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of ICCV*, pages 2980–2988.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2357–2368.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992.

Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, Sara Mahdavi, Jason Wei, Hyung Won Chung, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Sarvesh Soni and Dina Demner-Fushman. 2026a. A dataset for addressing patient's information needs related to clinical course of hospitalization. *Scientific Data*.

Sarvesh Soni and Dina Demner-Fushman. 2026b. Overview of the archehr-qa 2026 shared task on grounded question answering from electronic health records. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, Palma, Mallorca (Spain). ELRA.

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.

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.

# A. Prompt for Evidence Identification

## A.1. Evidence Identification Prompt (Stage 1)

We use large language model to identify sentences that are essential for answering the clinician question. The model performs binary classification for each sentence: *essential* or *not-essential*. The prompt template is shown below.

```
Task:
For each sentence, decide whether it is ESSENTIAL to answer the clinician
question.

Definitions:
- essential: directly necessary to answer the question.
- not-essential: not directly required.

Precision rule:
Only label as "essential" if removing the sentence would make answering
incomplete or incorrect.

Rules:
- Think step-by-step internally but DO NOT output reasoning.
- Output JSON only.
- Include ALL sentence ids.
- Allowed labels: "essential" or "not-essential"

===== FEW-SHOT EXAMPLES =====

Patient Narrative:
{PATIENT_NARRATIVE_EXAMPLE}

Clinician Question:
{CLINICIAN_QUESTION_EXAMPLE}

Sentences:
1: {SENTENCE_1}
2: {SENTENCE_2}
3: {SENTENCE_3}
...

Correct Output:
{
  "1": "essential",
  "2": "not-essential",
  "3": "essential"
}

===== END OF FEW-SHOT =====

Now classify:

Patient Narrative:
{PATIENT_NARRATIVE}

Clinician Question:
{CLINICIAN_QUESTION}

Sentences:
1: {SENTENCE_1}
```

```
2: {SENTENCE_2}
3: {SENTENCE_3}
...

Output JSON only.
```

### A.2. Self-Reflection Prompt (Stage 2)

After the initial prediction, we perform a second-pass self-reflection step where the model reviews its previous decisions and corrects potential mistakes.

```
You previously classified the sentences as:

{INITIAL_PREDICTION_JSON}

Now carefully review your previous classification.

Check:
1. Is any truly essential sentence missing?
2. Is any labeled "essential" actually background?
3. Are essential sentences strictly necessary?

Revise if needed.

Rules:
- Think step-by-step internally but DO NOT output reasoning.
- Output JSON only.
- Include ALL sentence ids.
- Allowed labels: "essential" or "not-essential"

Patient Narrative:
{PATIENT_NARRATIVE}

Clinician Question:
{CLINICIAN_QUESTION}

Sentences:
1: {SENTENCE_1}
2: {SENTENCE_2}
3: {SENTENCE_3}
...

Output JSON only.
```

## B. Prompt for Answer Generation

### B.1. Prompt for Clinical Question Intent Classification

We use large language model to classify the clinical intent of each question. The exact prompt used for intent classification is shown below.

```
You must classify the primary intent of this clinical QA case.

Choose one primary intent from:

1. procedure_justification
2. icu_or_monitoring_necessity
3. treatment_plan
4. medication_safety
5. pathophysiology_explanation
```

```
6. malignancy_concern
7. lab_abnormal_concern
8. recovery_duration
9. prognosis_severity
10. activity_clearance
11. persistent_unexplained_symptoms
12. neurologic_mental_status_change

Also optionally choose a secondary intent (or null).

Also assign a risk level:
low / moderate / high

Definitions:
- activity_clearance = travel, flying, oxygen removal, return to work
- malignancy_concern = fear of cancer
- pathophysiology_explanation = asking why X caused Y
- lab_abnormal_concern = abnormal lab value severity
- recovery_duration = asking how long recovery takes
- prognosis_severity = asking how serious / survival outlook
- medication_safety = tapering, toxicity, side effects
- persistent_unexplained_symptoms = chronic unexplained symptoms
- neurologic_mental_status_change = confusion, dementia, seizures
- treatment_plan = what should be done next
- procedure_justification = why surgery/procedure done
- icu_or_monitoring_necessity = why ICU or monitoring required

Patient Narrative:
{PATIENT_NARRATIVE}

Clinician Question:
{CLINICIAN_QUESTION}

Output JSON format:

{
  "primary": "...",
  "secondary": "...",
  "risk_level": "..."
}
```

### B.2. Answer Generation Prompt

```
You are a senior clinical specialist.

Your task is to generate a precise, evidence-grounded answer to the
Clinician Question using ONLY the provided Clinical Note.

PRIMARY OPTIMIZATION GOALS:

A) Maximize lexical overlap with the Clinical Note:
Reuse the exact medical terms and phrasing from the note whenever possible.
Prefer copying short clauses rather than paraphrasing.

B) Maximize clinically meaningful concept coverage:
Explicitly mention all key clinical concepts present in the note, including:
diagnoses, symptoms, procedures, imaging or laboratory tests, medications,
complications, anatomy, and clinically relevant numbers or timelines.
```

```
STRICT RULES:

- Use ONLY information documented in the Clinical Note.
- Do NOT add external knowledge.
- Do NOT speculate beyond the note.
- Prioritize essential sentences.
- Do NOT include sentence numbers, brackets, or citations.
- Use explicit clinical nouns rather than vague pronouns.
- Write in concise professional clinical register.
- Limit to 75 words maximum.
- Use 2-4 sentences.
- Avoid filler phrases.

COMPRESSION STRATEGY:

Combine related clinical facts into high-information sentences.
Retain specific terminology such as procedure names, medication names,
and diagnoses.
Preserve original phrasing where possible to increase textual alignment.

===== FEW-SHOT EXAMPLES =====

Patient Question:
{PATIENT_QUESTION_EXAMPLE}

Clinician Question:
{CLINICIAN_QUESTION_EXAMPLE}

Example Clinical Note:
1: {SENTENCE_1}
2: {SENTENCE_2}
3: {SENTENCE_3}

Example Relevant Sentences: [1, 3]

Answer:
{REFERENCE_ANSWER}

===== END OF FEW-SHOT =====

Now generate the answer.

Patient Question:
{PATIENT_QUESTION}

Clinician Question:
{CLINICIAN_QUESTION}

Clinical Note:
1: {SENTENCE_1}
2: {SENTENCE_2}
3: {SENTENCE_3}

Essential Sentences:
[ID_LIST]

Final Answer (<=75 words):
```