tt501 at ArchEHR-QA 2026: Few-Shot Prompting with Retrieval-Augmented Generation for Grounded Clinical EHR Question Answering

Tai Tran Tan
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
22521287@gm.uit.edu.vn

Abstract

We present the system developed by team tt501 for the ArchEHR-QA 2026 shared task, participating in three subtasks: Evidence Identification (Subtask 2), Answer Generation (Subtask 3), and Evidence Alignment (Subtask 4). All approaches rely exclusively on prompt engineering with reasoning-capable large language models from xAI (`grok-4-fast-reasoning` and `grok-4-1-fast-reasoning`) via API, requiring no task-specific fine-tuning. For Subtask 2, we propose a two-stage hybrid pipeline combining BM25 keyword retrieval with full-context Chain-of-Thought prompting and a secondary LLM refinement step, achieving a best Strict Micro-F1 of 58.84. For Subtask 3, we implement a retrieval-augmented generation (RAG) approach that chains predicted evidence from Subtask 2 with few-shot prompting to produce grounded 75-word clinical answers, reaching an overall score of 31.36. For Subtask 4, we develop a recall-optimised few-shot prompting strategy with reasoning annotations embedded in exemplars, enabling the model to correctly identify many-to-many evidence–answer alignments and achieving a best Micro-F1 of 79.11. Our results demonstrate that carefully engineered prompts with reasoning-capable LLMs can achieve competitive performance on fine-grained clinical evidence tasks without any task-specific training.

Keywords: clinical NLP, EHR question answering, evidence identification, retrieval-augmented generation, few-shot prompting, evidence alignment

1. Introduction

Responding to patient-generated messages through electronic health record (EHR) portals places a growing burden on clinical staff (Sinsky et al., 2016). A system capable of automatically answering such questions must do more than produce a fluent response: it must locate the relevant clinical evidence within the patient's own notes and present an answer that is directly grounded in that evidence. This requirement makes the problem inherently multi-step, touching on information retrieval, faithful text generation, and fine-grained citation.

The ArchEHR-QA 2026 shared task (Soni and Demner-Fushman, 2026b) decomposes this challenge into four subtasks, extending the 2025 edition that attracted 29 participating teams (Soni et al., 2025). We submitted systems for three of the four subtasks, forming a natural end-to-end pipeline. Subtask 2 asks systems to select the minimal set of note sentences that provide clinical evidence for a given patient question. Subtask 3 uses that evidence to generate a concise, grounded answer. Subtask 4 then requires attributing each sentence in the answer back to the supporting note sentences, producing an explicit citation structure.

Our approach is deliberately lightweight: all three subtasks are addressed through prompt engineering alone, using the `grok-4-1-fast-reasoning` model (xAI, 2025) via the xAI API, without any fine-tuning or supplementary training data. The central question we investigate is how much can be achieved through careful prompt design when the underlying model already has strong reasoning capabilities. This paper describes our system and discusses what worked and why across each subtask.

2. Task and Data

The ArchEHR-QA 2026 dataset (Soni and Demner-Fushman, 2026a) consists of 167 patient cases derived from the MIMIC database (Johnson et al., 2016). Each case includes a patient-authored question, a clinician-interpreted reformulation of that question (at most 15 words), a clinical note excerpt with sentence-level relevance labels (*essential*, *supplementary*, or *not-relevant*), a reference clinician-authored answer, and gold evidence–answer alignment annotations. The dataset is divided into a development set (cases 1–20) and two test splits (cases 21–120 and 121–167).

The shared task defines four subtasks. We participate in the three that form a coherent retrieval-generation-attribution pipeline and do not submit to Subtask 1 (question interpretation). The three subtasks addressed in this work are:

Subtask 2 (Evidence Identification) Given the note and the patient question, systems must predict the minimal set of note sentences sufficient to

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 497–505
12 May 2026.

answer the question. Performance is measured by Precision, Recall, and F1 over predicted sentence-ID sets under two variants: *Strict* (essential sentences only) and *Lenient* (essential and supplementary sentences combined). The official test split covers 47 cases (IDs 121–167).

Subtask 3 (Answer Generation) Using the identified evidence, systems must generate a response of at most 75 words that is directly grounded in the note. Answers are evaluated against reference clinician responses using BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2020), AlignScore (Zha et al., 2023), and MEDCON (Ben Abacha et al., 2023); the overall score is the arithmetic mean of all six. The evaluation covers the same 47 cases as Subtask 2.

Subtask 4 (Evidence Alignment) Given a generated answer and the associated note, systems must align each answer sentence to the note sentence(s) from which it derives, producing a many-to-many mapping. Performance is measured by Micro and Macro Precision, Recall, and F1 over predicted sentence-pair links. The evaluation covers 147 cases (IDs 21–167).

3. System Description

All three subtasks use the xAI API with temperature set to 0 for deterministic outputs and JSON-structured responses parsed with regex fallbacks for robustness.[1] Two model versions were used: `grok-4-fast-reasoning` for earlier runs and `grok-4-1-fast-reasoning` (xAI, 2025) for subsequent runs, both reasoning-capable large language models from xAI. For each subtask we submitted two runs, designed so that the second run directly addresses a known limitation of the first. All scripts and prompt templates are publicly available at urlhttps://github.com/taitran501/tt501-archehr-qa-2026.

3.1. Subtask 2: Evidence Identification

Selecting the right evidence sentences requires balancing recall against precision: missing a relevant sentence is just as harmful as including an irrelevant one. We explored two strategies that differ primarily in how much of the note is exposed to the language model.

Run 1: Hybrid BM25 and LLM Reranker (*hybrid-bm25-llm*). The first run uses BM25 (Robertson and Zaragoza, 2009) as a first-stage retriever, selecting the top-20 candidate sentences per case using the clinician-interpreted question as the query. These candidates are then passed to `grok-4-fast-reasoning` in a single API call with a precision-focused prompt that instructs the model to return only the sentence IDs that directly and sufficiently answer the question. This approach is computationally efficient, processing 47 cases in approximately three minutes, but it inherits BM25's sensitivity to vocabulary mismatch: evidence expressed in clinical terminology that does not appear verbatim in the question can be filtered out before the LLM ever sees it.

Run 2: Full-Context Chain-of-Thought Ensemble with Refinement (*ensemble-refine-v1*). The second run presents the complete note excerpt to the LLM, removing the first-stage filter entirely. The prompt is oriented toward recall, instructing the model to trace the full clinical narrative (problem, intervention, complication, and resolution) and to include all causally relevant sentences. Two annotated development examples serve as few-shot demonstrations. A secondary LLM call then reviews the initial selection, filtering out noise and resolving cases where patient phrasing and clinical documentation use different terminology for the same concept.

3.2. Subtask 3: Answer Generation

Subtask 3 requires generating a professional, grounded answer of at most 75 words, without introducing clinical claims that are not supported by the note.

Run 1: Zero-Shot Full-Note Generation (*zero-shot-v1*). The first run feeds the entire note excerpt to `grok-4-fast-reasoning` without any pre-selected evidence or few-shot examples. The system prompt specifies the 75-word limit, the requirement for a professional medical register, and strict grounding to the note. Output is requested as a JSON object with an `answer` field to prevent formatting artefacts.

Run 2: Retrieval-Augmented Generation with Few-Shot Examples (*rag-fewshot-v1*). The second run implements a retrieval-augmented generation pipeline that chains directly from the Subtask 2

[1] The xAI API returns JSON reliably at temperature 0; regex fallback was triggered in fewer than 3% of responses, exclusively due to occasional leading/trailing whitespace or markdown code-fence artefacts wrapping an otherwise valid JSON object. The fallback strips these wrappers and re-parses; if parsing still fails the case is logged and excluded rather than silently producing incorrect output.

output. The sentences predicted as evidence by *ensemble-refine-v1* are resolved to their text and presented as a structured block, reducing the input to only the sentences considered relevant. Three gold-annotated development cases are included as few-shot demonstrations, each pairing the patient question, the evidence sentences, and the reference clinician answer. This design encourages the model to produce responses that are lexically and semantically close to how a clinician would write, and the reduced context appears to help calibrate response length.

3.3. Subtask 4: Evidence Alignment

Evidence alignment requires mapping each answer sentence to the note sentence(s) from which it was derived. The task is many-to-many: a single answer sentence may synthesise information from several parts of the note, and under-citing is a common failure mode when a prompt only rewards identifying the most lexically similar match.

Run 1: Basic Few-Shot Alignment (*grok-4-1-reasoning-fewshot-v1*). The first run supplies the clinician-interpreted question, the note excerpt, and the answer sentences to the model. A few-shot prompt with development examples illustrates the expected JSON output format consisting of `{answer_id, evidence_id}` objects. This provides a clean baseline for how much the model can align without explicit guidance on *how* to reason about citations.

Run 2: Recall-Optimised Few-Shot with Reasoning Annotations (*few-shot-reasoning-v2*). The second run introduces two changes over Run 1. First, the patient-authored question is added alongside the clinician question, because the patient's own wording often mirrors the lay language used in the answer and can serve as a bridge to clinical terminology in the note. Second, and more importantly, each of the four few-shot examples is augmented with explicit reasoning: for each citation link, the prompt explains in plain language why that note sentence supports that answer sentence. This teaches the model the rationale for citing, not only the format. The system prompt also includes an explicit instruction that answer sentences frequently synthesise information from multiple note sentences, counteracting the natural tendency to select only the most lexically similar source.

4. Results and Analysis

Tables 1–3 summarise official test-set scores for each submitted run. Run 1 of Subtasks 2 and 3 uses `grok-4-fast-reasoning`; all other runs use `grok-4-1-fast-reasoning`. All experiments run with temperature 0 and no external resources beyond the shared-task data.

Run	Strict			Lenient		
	P	R	F1	P	R	F1
hybrid-bm25-llm	50.1	56.0	52.9	66.7	56.0	60.9
ensemble-refine-v1	56.0	61.9	**58.8**	74.9	61.9	**67.8**

Table 1: Subtask 2 test results (IDs 121–167, n=47). Micro Precision (P), Recall (R), F1 (%).

Run	BLEU	R-L	SARI	BERT	MED	ALN	Ovrl
zero-shot-v1	3.0	19.0	53.5	36.8	36.6	19.4	28.1
rag-fewshot-v1	6.2	24.0	58.3	41.4	35.9	22.4	**31.4**

Table 2: Subtask 3 test results (n=47). All scores in %. BERT=BERTScore, MED=MEDCON, ALN=AlignScore, Ovrl=overall (mean of all six).

4.1. Subtask 2: Evidence Identification

The Subtask 2 results highlight the value of exposing the model to the full clinical note rather than a BM25-filtered subset. With complete context and a Chain-of-Thought prompt, the ensemble run (Run 2) achieves a Strict Micro-F1 of 58.8 and a Lenient F1 of 67.8 (Table 1), indicating that the model can reliably identify sentences that carry the core clinical evidence. Its Strict recall of 61.9 suggests that many relevant sentences are recovered even when they are only indirectly related to the surface wording of the question, a setting in which lexical retrieval alone is brittle. The consistent gap between Lenient and Strict F1 across runs further points to the presence of sentences that provide background or clarifying information: they are not strictly necessary for correctness, but they enrich the clinical story presented to the patient.

We note that the two runs also differ in model version: Run 1 uses `grok-4-fast-reasoning`, whereas Run 2 uses `grok-4-1-fast-reasoning`. The observed improvement in Strict F1 from 52.9 to 58.8 therefore reflects the combined effect of the switch to full-context prompting *and* the newer model version, and cannot be attributed exclusively to prompt design. Isolating the contribution of each factor would require rerunning both prompts under the same model version, which we leave for future work.

4.2. Subtask 3: Answer Generation

In Subtask 3, the contrast between the two runs reflects the benefit of separating retrieval from generation and showing the model concrete clinical examples. When the model receives only the full note and a zero-shot prompt, it must jointly decide which parts of the note matter and how to phrase

Run	Micro			Macro		
	P	R	F1	P	R	F1
fewshot-v1	85.1	72.5	78.3	86.4	75.6	78.5
fewshot-v2	78.2	80.1	**79.1**	80.0	82.6	**79.7**

Table 3: Subtask 4 test results (IDs 21–167, n=147). Precision (P), Recall (R), F1 (%). *fewshot-v1*: basic few-shot, clinician question only. *fewshot-v2*: recall-optimised, reasoning annotations, patient+clinician question.

the answer. Conditioning instead on sentences already identified as evidence, together with three few-shot demonstrations, leads to answers that better align with the reference along lexical, semantic, and factual dimensions: the RAG few-shot run (Run 2) attains an overall score of 31.4 compared with 28.1 for the zero-shot baseline, with consistent improvements across BLEU, ROUGE-L, SARI, BERTScore, MEDCON, and AlignScore (Table 2). The behaviour around the 75-word constraint reinforces this picture: 16 zero-shot predictions require truncation, whereas only a single RAG prediction does so, suggesting that the demonstrations also help the model internalise an appropriate answer length and professional register.

As with Subtask 2, the two runs differ in model version: Run 1 uses `grok-4-fast-reasoning` and Run 2 uses `grok-4-1-fast-reasoning`. Consequently, the gain in overall score from 28.1 to 31.4 reflects both the switch to a RAG pipeline with few-shot examples and the use of an updated model, and neither factor can be fully isolated from the other.

4.3. Subtask 4: Evidence Alignment

For evidence alignment (Table 3), the recall-optimised prompt with reasoning annotations encourages the model to treat citation as a genuinely many-to-many mapping rather than a one-to-one matching problem. The corresponding run (Run 2) reaches a Micro-F1 of 79.1, with recall of 80.1 and precision of 78.2, while the basic few-shot baseline (Run 1) attains a Micro-F1 of 78.3 with recall and precision of 72.5 and 85.1. This pattern is consistent with the prompt's instruction to cite *all* sentences that contribute to a claim: some precision is traded for recall, but the overall balance better reflects the evaluation protocol, which penalises both missing and spurious links. The reasoning annotations in the exemplars make this behaviour concrete by spelling out how each citation connects an answer span to specific parts of the note, and the model appears to emulate this strategy at test time. Adding the patient-authored question to the context further helps bridge lay phrasing in the answer and clinical terminology in the note, making it easier for the model to recognise when multiple note sentences jointly support a single statement.

4.4. Qualitative Error Analysis

To clarify where each approach succeeds and remains fragile, we examine representative failure cases drawn from the development set.

Subtask 2: vocabulary mismatch in BM25 retrieval. The hybrid run (Run 1) uses BM25 with the clinician question as query, filtering the note excerpt to a top-20 candidate set before the LLM sees any evidence. In Case 4 (cardiac catheterisation), the clinician question asks *"Why was cardiac catheterization recommended?"* One gold-essential sentence reads: *"Last echo showed LVEF = 25%"*, the key clinical finding that motivated the procedure, but the abbreviations *LVEF* and *echo* do not appear in the question. BM25 assigns it a low retrieval score, and the sentence is below the top-20 cut-off, so the LLM reranker never sees it. The ensemble run (Run 2) presents the full note, and the CoT prompt's instruction to trace the complete clinical narrative (*"What led to this decision?"*) correctly recovers the sentence. This pattern (essential context buried in clinical abbreviations not shared with the patient-facing question) recurs across cardiovascular and ICU cases and accounts for a substantial fraction of Run 1's false negatives.

Subtask 3: incomplete grounding in zero-shot generation. The zero-shot run (Run 1) is prone to generating answers that cover only the first relevant thread in the evidence, ignoring subsequent instructions or alternative findings. In Case 3 (traumatic brain injury discharge instructions), the gold answer spans five sentences: expected symptoms (drowsiness, headaches, dizziness, irritability, memory loss), a prognosis note (symptom decrease over weeks), a follow-up instruction (TBI specialist referral), a general recovery reminder, and a list of warning signs requiring urgent re-contact (visual changes, unilateral weakness, speech difficulty). Run 1 produced a single sentence: *"Symptoms such as irritability and headaches are normal after a traumatic brain injury and should decrease over the next several weeks"* – correct but radically incomplete. The RAG few-shot run (Run 2) mitigates this by conditioning on pre-selected evidence sentences and providing three complete reference answers as exemplars, which anchor the model's expected output length and content breadth.

Subtask 4: under-citation of synthesised answer sentences. The basic few-shot baseline (Run 1) confidently assigns a single note sentence to each answer sentence, producing high precision

(85.1) at the cost of recall (72.5). In Case 5 (musculoskeletal chest pain), the first answer sentence *"The chest pain was musculoskeletal, not overdose or cardiac related"* is a conclusion synthesised from four separate note sentences: the musculoskeletal exam finding, the normal EKG, the negative cardiac enzymes, and the normal transthoracic echocardiogram. Run 1 cites only the exam finding sentence (`[9]`). The recall-optimised run (Run 2) includes reasoning annotations in its exemplars that explicitly model this synthesis pattern (e.g., *"Answer mentions no cardiac causes: cite EKG sentence [10], enzyme sentence [18], and echo sentence [19]"*), teaching the model that a conclusory answer sentence can require multi-source support, and thereby recovering the missing citations.

4.5. Inference Cost and Latency

Both `grok-4-fast-reasoning` and `grok-4-1-fast-reasoning` share the same pricing tier on the xAI API: $0.20 per million input tokens and $0.50 per million output tokens (reasoning tokens billed at the output rate). Submitting via the Batch API reduces all token costs by 50% relative to synchronous calls.

Table 4 summarises the API usage pattern for each submitted run. Subtask 2 Run 1 processes all 47 cases via synchronous API calls with a BM25 pre-filter, completing in approximately three minutes. All other runs use the xAI async Batch API, submitting all cases in a single batch job and polling for results, benefiting from the 50% batch discount. Batch jobs for the 47-case Subtasks 2–3 sets typically completed within approximately 10 minutes. The 147-case Subtask 4 batches required approximately 15 minutes, reflecting the larger per-case input context. Input contexts for the full-context runs (Subtask 2 Run 2 and both Subtask 3 runs) are substantially larger than those of the BM25-filtered run because the entire note excerpt rather than a top-20 candidate subset is provided to the model.

Despite the reasoning model overhead and the full-context prompts, the combined cost of each individual run remained below $0.20 USD, and the total spend across all six test-set runs was under $1.00. This demonstrates that prompt-engineering approaches on small shared-task datasets are viable even on a personal API budget.

Subtask	Run	Cases	API Mode	Time	Cost
2	hybrid-bm25-llm	47	Synchronous	∼3 min	<$0.20
2	ensemble-refine-v1	47	Async Batch	∼10 min	<$0.20
3	zero-shot-v1	47	Async Batch	∼10 min	<$0.20
3	rag-fewshot-v1	47	Async Batch	∼10 min	<$0.20
4	fewshot-v1	147	Async Batch	∼15 min	<$0.20
4	fewshot-v2	147	Async Batch	∼15 min	<$0.20

Table 4: Inference cost and latency per run on the official test sets. Wall-clock time is approximate and subject to xAI server load. Cost is based on the xAI pricing at competition time: $0.20/$0.50 per million input/output tokens (standard); Batch API runs benefit from a 50% discount. Total spend across all six runs was under $1.00.

5. Conclusion

This paper described a prompt-only system for three subtasks of the ArchEHR-QA 2026 shared task. Working entirely through prompt engineering on a single reasoning-capable language model, without any task-specific fine-tuning, we found that the design of the prompt itself has a substantial effect on performance in each subtask.

For evidence identification, the key insight is that restricting the model's input to BM25-selected candidates limits recall in ways that cannot be recovered by downstream reranking. Allowing the model to read the full note and reason over the complete clinical narrative is more effective, even at a higher inference cost. For answer generation, few-shot demonstrations drawn from annotated development cases help the model adopt the register and length expected by clinical evaluators, and coupling generation with retrieved evidence produces answers that are more faithful to the source note. For evidence alignment, the most impactful change was adding explicit reasoning to the few-shot exemplars: rather than showing the model what correct citations look like, explaining why each citation is correct guides the model to reason about multi-source attribution rather than defaulting to surface lexical similarity.

Taken together, these results support the broader observation that reasoning-capable language models can be effective on structured clinical NLP tasks when prompts are designed to make the reasoning process explicit, rather than leaving the model to infer it from output examples alone. Future work will examine joint optimisation across the subtask pipeline, self-consistency strategies for evidence identification, and principled methods for selecting few-shot demonstrations.

6. Limitations

Reproducibility. All results depend on the `grok-4-1-fast-reasoning` and `grok-4-fast-reasoning` APIs provided by xAI. Model outputs may change as providers update or deprecate model versions, reducing exact reproducibility over time. Temperature is set to 0 throughout to mitigate non-determinism, but API-level changes remain outside our control.

Model version confound. Run 1 of Subtasks 2 and 3 uses `grok-4-fast-reasoning`, whereas all other runs use `grok-4-1-fast-reasoning`. The observed improvements in Subtask 2 (Strict F1: 52.9 $\rightarrow$ 58.8) and Subtask 3 (overall: 28.1 $\rightarrow$ 31.4) therefore reflect the combined effect of prompt and pipeline changes *and* a model version upgrade; the contribution of each factor cannot be fully isolated without rerunning all configurations under the same model, which we leave for future work. The mixed versioning was intentional: `grok-4-fast-reasoning` was the model available when Subtasks 2 and 3 development began. As experimentation progressed it was superseded by `grok-4-1-fast-reasoning`, which xAI positioned as the recommended reasoning successor; Subtask 4 was designed from the outset to use `grok-4-1-fast-reasoning` for both runs, so performance differences there reflect only prompt and pipeline changes. As of March 2026, `grok-4-fast-reasoning` is no longer listed on the active xAI models page; researchers replicating this work should use `grok-4-1-fast-reasoning` or its current successor.

Error propagation in the RAG pipeline. Subtask 3 results are conditioned on Subtask 2 predictions. Errors in evidence identification directly degrade the quality of generated answers. An analysis of cases where evidence sets were incomplete or incorrect is left for future work.

No human evaluation. Our system is assessed solely through automatic metrics. For clinical applications, human evaluation by domain experts would be necessary to verify factual accuracy, appropriate professional register, and patient safety. This limitation is particularly consequential for Subtask 3, where the generated answer is presented directly to a patient. Automatic metrics such as BLEU, ROUGE-L, and BERTScore measure lexical and semantic overlap with a reference answer but cannot detect clinically unsafe content, for example an omitted contraindication, a misattributed finding, or phrasing that could be misinterpreted in a self-care context. Human review by a clinician expert would be required before any such system could be considered for deployment in a patient-facing setting.

Dataset size. The ArchEHR-QA 2026 test set contains 47 cases for Subtasks 2–3 and 147 cases for Subtask 4. Score differences of <1 point should be interpreted with caution given the limited sample sizes.

7. Ethical Considerations

The ArchEHR-QA dataset is a de-identified derivative of the MIMIC database (Johnson et al., 2016), distributed by the shared-task organisers for competition use. All clinical data were de-identified prior to inclusion in the released corpus; no fields that could re-identify individuals are present.

During experimentation, this data was submitted to the xAI API under an individual account. We note that the xAI consumer Terms of Service permit use of inputs for model improvement by default, and the PhysioNet Data Use Agreement recommends verifying zero-retention guarantees before transmitting MIMIC-derived data to any third-party service. We acknowledge this as an unresolved compliance concern. Researchers replicating this work are strongly advised to use an enterprise API agreement with explicit zero-retention and no-training guarantees, or a locally hosted model, and to independently verify alignment with the PhysioNet DUA applicable to their own credentials.

This work does not introduce new patient data, does not perform any re-identification attempts, and does not deploy the system in a clinical setting. The system is intended solely for research purposes and should not be used for clinical decision-making without further validation.

8. Bibliographical References

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An investigation of evaluation metrics for automated medical note generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 15128–15143.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

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 of the Association for Computational Linguistics (ACL 2002)*, pages 311–318.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760.

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)*. ELRA.

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 BioNLP Workshop at ACL 2025*.

xAI. 2025. Grok-4: Advancing reasoning and scientific discovery. https://x.ai/blog/grok-4. Accessed 2026.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 401–415.

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 (ACL 2023)*, pages 11328–11348.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

A. Prompts and Few-Shot Examples

This appendix provides the exact system prompts, user prompt templates, and few-shot examples used for each submitted run. Variable placeholders are shown in `{curly_braces}`. All runs use `temperature=0` and JSON-structured output.

Subtask 2, Run 1: *hybrid-bm25-llm*

The BM25 first stage selects the top-20 candidates using the clinician question as the query. The LLM reranker then receives the following prompt, where `{clinician_question}` is the question text and the note is limited to the BM25-selected sentences.

System message.

> *You are a clinical documentation specialist.*

User message.

> **Question:** {clinician_question}
>
> **Task:** Identify all sentences in the Note Excerpt below that provide evidence answering this question.
>
> **Output Format:** Respond ONLY with a JSON object: `{"evidence_ids": ["0", "2", ...]}`

Subtask 2, Run 2: *ensemble-refine-v1*

The full note excerpt is provided to the model. The prompt is recall-oriented and uses two annotated development examples as few-shot demonstrations.

System message.

> *You are a precise medical evidence classifier. Respond ONLY with valid JSON.*

User message template.

> You are an expert medical documentation specialist. Your task is to identify the COMPREHENSIVE set of sentences from the clinical note that provide EVIDENCE to answer the clinician's question.
>
> **Instructions:**
>
> 1. Read the **Clinician Question** and the **Clinical Note** carefully.
> 2. Select **ALL** sentences that: (a) provide direct answers, (b) provide necessary context (timeline, initial conditions, reasons for changes), (c) explain WHY a decision was made or HOW an outcome occurred, (d) detail specific values, dates, or test results relevant to the inquiry.
> 3. Completeness is critical. It is better to include a borderline sentence than to miss it.
> 4. Capture the full narrative arc: Problem → Intervention → Complication → Resolution.

Few-Shot Examples:

Example 1
Question: Why was ERCP recommended over medication for CBD sludge?
Note: [1] During the ERCP a pancreatic stent was required... [7] As the patient's INR was normalized to 1.2, a sphincterotomy was safely performed, with removal of several biliary stones...
Answer: `{"relevant_ids": ["1", "5", "6", "7"]}`

Example 2
Question: Why was the patient given Lasix and oxygen reduced?
Note: [1] Acute diastolic heart failure: Pt developed signs and symptoms of volume overload... [6] This undoubtedly contributed to his continued hypoxemic respiratory failure.
Answer: `{"relevant_ids": ["1", "4", "5", "6"]}`

Current Task:
Clinician Question: {question}
Clinical Note:
{note_text}

Analyze step-by-step:

1. Identify the core inquiry.
2. Trace the relevant storyline (chronological, cause-and-effect).
3. Select ALL sentence IDs that form the complete evidence chain.

Output Format: Respond ONLY with valid JSON: `{"relevant_ids": ["1", "3", "5"]}`

Subtask 3, Run 1: *zero-shot-v1*

The entire note excerpt is provided; no pre-selected evidence or few-shot examples are used.

System message.

> *You are a clinical documentation specialist extracting factual answers from medical notes.*
>
> Critical Guidelines:
>
> 1. Extract ONLY information explicitly stated in the note—no speculation, inference, or hedging.
> 2. Use professional medical register ("The patient..."). No empathy phrases.
> 3. 75 words maximum (strictly enforced).
> 4. Use past tense for completed medical events.
> 5. State facts directly without qualifiers.

User message template.

> **Patient Question:** {patient_question}
> **Clinician Question:** {clinician_question}
> **Note Excerpt:** {full_note_text}
>
> **Output Format:** Respond ONLY with valid JSON: `{"answer": "Your factual extraction here (max 60 words)"}`

Subtask 3, Run 2: *rag-fewshot-v1*

Evidence sentences predicted by *ensemble-refine-v1* are resolved to their text and passed instead of the full note. Three development-set reference answers serve as few-shot demonstrations.

System message.

> *You are a professional medical scribe. Your task is to draft a concise clinical answer to a clinician's question, based ONLY on the provided evidence sentences.*
>
> Constraints:
>
> 1. **Length:** Maximum 75 words (≈5 sentences).
> 2. **Tone:** Professional, detached, medical register ("The patient..."). Do NOT use "You" or conversational language.
> 3. **Grounding:** Use ONLY the provided evidence. Do not hallucinate or use external knowledge.
> 4. **Format:** Plain text paragraph. No citations. No markdown.

User message template (few-shot prefix). Three examples taken from the development set are prepended (abbreviated here for space). The full examples correspond to Cases 4, 1, and 9 from the ArchEHR-QA 2026 development data.

> **Example Case:**
> **Question:** Why was cardiac catheterization recommended to the patient?
> **Evidence:**
> [selected essential sentences from Case 4]
> **Answer:** The patient was recommended a cardiac catheterization for worsening heart failure confirmed by left ventricle ejection fraction of 25% on his echocardiogram...
>
> ⟨two further examples omitted for brevity⟩
>
> **Current Case:**
> **Question:** {clinician_question}
> **Evidence:**
> {evidence_text}
> **Answer:**

Subtask 4, Run 1: *fewshot-v1*

Two development-set examples are used. The patient question is omitted from context. No reasoning annotations are included in the exemplars.

System message.

You are an expert clinical evidence aligner. Your task is to align each answer sentence with the supporting sentence(s) from the clinical note excerpt. For each answer sentence (identified by Answer ID), return a JSON object containing the Answer ID and an array of Evidence IDs from the Note Sentences that support it. If an answer sentence is not supported by any note sentence, return an empty array [] for its evidence_id.

Constraints:

1. Return ONLY valid JSON as a list of objects. Do NOT wrap in Markdown.
2. Format: `[{"answer_id": "1", "evidence_id": ["2","5"]}, ...]`
3. Avoid over-citing.
4. Alignments are many-to-many.

User message template.

Example Case:
Clinician Question: {clinician_question}
Note Excerpt:
[1]... [N]...
Answer Sentences:
[Answer 1]...
Output: `[ {"answer_id":"1", "evidence_id":["2"]}, ...]`

⟨second example follows the same pattern⟩

Current Case:
Clinician Question: {clinician_question}
Note Excerpt:
{note_text}
Answer Sentences:
{answer_text}
Output:

Subtask 4, Run 2: *fewshot-v2*

Four development-set examples (Cases 1, 2, 5, 6) are used. The patient question is added to context. Exemplars include reasoning annotations explaining why each note sentence supports the corresponding answer sentence.

System message.

You are an expert clinical evidence aligner specialised in grounding answer sentences to clinical note excerpts.

Key Alignment Principles:

1. An answer sentence often synthesises information from MULTIPLE note sentences. When this happens, cite ALL contributing note sentences.
2. A note sentence supports an answer sentence if it provides specific clinical facts referenced in the answer—even if the wording differs (paraphrasing is common).
3. If an answer sentence is a general conclusion NOT grounded in any specific note sentence, assign an empty array [].
4. Do NOT over-cite: only cite note sentences whose information is actually referenced in the answer sentence.

Output Rules:

1. Return ONLY valid JSON. NO Markdown code fences.
2. Format: `[ {"answer_id":"1", "evidence_id":["2","5"]}, ...]`
3. Include an entry for EVERY answer sentence in order.
4. Alignments are many-to-many.

User message template (abbreviated few-shot prefix).

Example Case (ID 1):
Patient Question: ...
Clinician Question: Why was ERCP recommended...
Note Excerpt:
[1]... [N]...
Answer Sentences:
[Answer 1]...
Reasoning:
Answer 1: "ERCP was recommended to place a CBD stent" → Note [2] describes ERCP placing a CBD stent.
...
Output: `[ {"answer_id":"1", "evidence_id":["2"]}, ...]`

⟨Examples for Cases 2, 5, and 6 follow the same pattern.⟩

Current Case:
Patient Question: {patient_question}
Clinician Question: {clinician_question}
Note Excerpt:
{note_text}
Answer Sentences:
{answer_text}
Output: