GigitAI at ArchEHR-QA 2026: Prompting Strategies and Constitutional AI for Clinical Question Answering

Saran Krishnasamy, Inez Wihardjo
GigitAI
San Francisco, CA
saran@gigit.ai, inez@gigit.ai

Abstract

Answering patient questions from electronic health records requires identifying relevant evidence in lengthy clinical notes and generating faithful, patient-friendly answers. We present a systematic study of LLM prompting strategies for both tasks, evaluating 21 evidence identification methods and 13 answer generation methods across 5 language models. For evidence identification, we find that LLM prompting outperforms traditional retrieval (BM25, SBERT, BioLinkBERT) by 19 F1 points, and that prompt framing alone controls precision/recall trade-offs: inclusive framing achieves 90% recall on dev while balanced framing reaches 67% precision. For answer generation, we apply a Constitutional AI pipeline that critiques and revises answers against five clinical faithfulness principles, improving BLEU and ROUGE over the constrained baseline. Our analysis reveals that chain-of-thought effectiveness is strongly model-dependent, and that simple well-designed prompts outperform complex multi-step pipelines. We evaluate our approaches on the ArchEHR-QA 2026 shared task at CL4Health, achieving 58.0 F1 for evidence identification and 31.8 overall for answer generation.

Keywords: clinical NLP, question answering, electronic health records, evidence identification, constitutional AI, large language models

1. Introduction

Patients discharged from hospitals frequently have unresolved questions about their diagnoses, treatments, and care plans (Singhal et al., 2023). Electronic health records (EHRs) contain detailed clinical information that could address these questions, but the complexity and volume of clinical notes make them inaccessible to most patients (Thirunavukarasu et al., 2023). Automatically answering patient questions from EHR data requires both identifying relevant evidence within lengthy clinical notes and generating accurate, faithful answers grounded in that evidence.

The ArchEHR-QA 2026 shared task (Soni and Demner-Fushman, 2026b) addresses this challenge through three subtasks: question interpretation (Subtask 1), evidence sentence identification (Subtask 2), and patient-friendly answer generation (Subtask 3). We participate in Subtasks 2 and 3 as team **GigitAI**.

Our approach centers on a comprehensive, systematic comparison of methods rather than a single-system submission. For Subtask 2, we evaluate 21 methods across 5 models, spanning traditional retrieval baselines and LLM prompting strategies. For Subtask 3, we evaluate 13 methods including a Constitutional AI pipeline. Our key contributions are:

1. A systematic comparison demonstrating that LLM prompting outperforms retrieval baselines by 19 F1 points for clinical evidence identification.
2. Analysis of prompt framing effects, showing that wording controls precision/recall trade-offs (inclusive: R=90.1, balanced: P=67.2).
3. A principle-guided critique-and-revise pipeline (adapted from Constitutional AI) with five clinical principles for answer generation, improving BLEU by +1.18 and ROUGE-Lsum by +1.67 over the constrained baseline.
4. Analysis of model-prompt interaction effects revealing that chain-of-thought is highly model-dependent (helps GPT-5 but catastrophically hurts Claude Haiku).

2. Related Work

2.1. Clinical QA from EHRs

Clinical question answering from EHRs has gained attention with datasets such as emrQA (Pampari et al., 2018) and EHRNoteQA (Kweon et al., 2024). The ArchEHR-QA dataset (Soni and Demner-Fushman, 2026a) established benchmarks for multi-turn QA from discharge summaries, with the 2025 shared task (Soni et al., 2025) revealing that LLM-based approaches generally outperformed retrieval-only methods for evidence identification. Recent work has demonstrated that large language models encode substantial clinical knowledge (Singhal et al., 2023), though grounding answers in specific patient records remains challenging.

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 565–577

2.2. Evidence Extraction and Retrieval

Traditional retrieval approaches for clinical evidence include sparse methods like BM25 (Robertson and Zaragoza, 2009) and dense retrieval with Sentence-BERT (Reimers and Gurevych, 2019). Domain-specific models such as BioLinkBERT (Yasunaga et al., 2022) and MedCPT (Jin et al., 2023) have shown improvements for biomedical text. Clinical-Longformer (Li et al., 2022) handles the long sequences typical of clinical notes, while cross-encoder reranking (Nogueira and Cho, 2019) with models like BGE (Xiao et al., 2024) provides more precise relevance scoring at higher computational cost.

2.3. LLM Prompting Strategies

Chain-of-thought prompting (Wei et al., 2022) has been shown to improve reasoning in LLMs, and self-consistency (Wang et al., 2023) further improves robustness by sampling multiple reasoning paths. DSPy (Khattab et al., 2024) and its MIPROv2 optimizer (Opsahl-Ong et al., 2024) enable automatic prompt optimization. Medical prompt engineering has been explored through approaches like Med-PaLM (Singhal et al., 2023) and Medprompt (Nori et al., 2023), which combine few-shot ensembling with chain-of-thought for clinical benchmarks. Bogireddy et al. (2025) achieved 2nd place at ArchEHR-QA 2025 using DSPy-optimized prompts with self-consistency voting.

2.4. Constitutional AI

Constitutional AI (Bai et al., 2022) was introduced as a training-time method for aligning AI systems to be helpful, harmless, and honest, using self-critique against explicit principles followed by reinforcement learning from AI feedback (RLAIF). While the original framework operates during model training, its core mechanism (evaluating outputs against a set of written principles and revising to resolve violations) can be applied at inference time as a prompting strategy. We adapt this principle-guided critique-and-revise paradigm for clinical answer generation, defining domain-specific principles focused on factual faithfulness, terminology accuracy, and temporal consistency with the source clinical notes. Our approach is also related to self-refinement prompting (Madaan et al., 2023), which iteratively improves outputs using self-generated feedback; we differentiate our pipeline by grounding the critique in domain-specific clinical principles rather than general quality criteria. Our use of the term "Constitutional AI" refers to this inference-time adaptation rather than the full RLAIF training procedure.

3. Task Description and Data

3.1. Subtask 2: Evidence Identification

Given a patient question, its clinical interpretation, and the full discharge summary segmented into numbered sentences, the task is to identify which sentences constitute essential or supplementary evidence for answering the question. Evaluation uses strict and lenient micro-averaged F1 scores, where lenient scoring gives partial credit for supplementary evidence.

3.2. Subtask 3: Answer Generation

Given the same inputs, generate a patient-friendly answer in at most 75 words. Evaluation combines six metrics: BLEU (Papineni et al., 2002), ROUGE-Lsum (Lin, 2004), SARI (Xu et al., 2016), BERTScore (Zhang* et al., 2020), AlignScore (Zha et al., 2023), and MEDCON (Yim et al., 2023), with the overall score being the average.

3.3. Dataset

The ArchEHR-QA 2026 dataset (Soni and Demner-Fushman, 2026a) consists of discharge summaries from MIMIC-IV (Johnson et al., 2024) with patient questions and clinician-authored answers. The data was accessed under a PhysioNet (Goldberger et al., 2000) credentialed data use agreement. Only de-identified discharge summaries as provided by the shared task organizers were used; MIMIC-IV data is already de-identified per HIPAA Safe Harbor standards. Three splits are available: a dev set (20 cases) with gold evidence labels, a test set released from the 2025 shared task (Soni et al., 2025) (100 cases with gold labels, used for additional development), and a new blind test-2026 set (47 cases) for official evaluation. Table 1 summarizes the dataset statistics.

	Dev	Test	Test-2026
Number of cases	20	100	47
Avg. sentences/note	21.4	26.0	34.0
Avg. evidence sentences	7.3	–	–

Table 1: Dataset statistics for ArchEHR-QA 2026. Test is from the 2025 edition (gold labels available); Test-2026 is the new blind evaluation set.

4. System Description

4.1. Subtask 2: Evidence Identification

We evaluate 21 methods organized into four categories (Figure 1): non-LLM retrieval baselines, LLM prompting variants, novel methods, and hybrid/ensemble approaches.

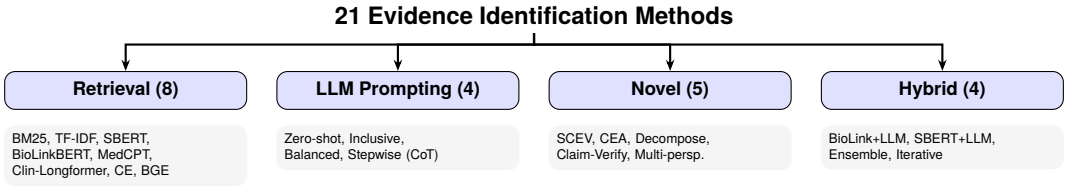

Figure 1: Overview of 21 evidence identification methods organized into four categories.

4.1.1. Non-LLM Retrieval Baselines

We implement eight retrieval baselines using the patient question and clinician interpretation as queries:
Sparse retrieval: BM25 (Robertson and Zaragoza, 2009) and TF-IDF (Manning et al., 2008) with cosine similarity.
Dense retrieval: Sentence-BERT (all-MiniLM-L6-v2) (Reimers and Gurevych, 2019), BioLinkBERT-large (Yasunaga et al., 2022), MedCPT (Jin et al., 2023), and Clinical-Longformer (Li et al., 2022).
Reranking: Cross-encoder reranking (MiniLM (Wang et al., 2020) bi-encoder + cross-encoder) and BGE-large retrieval with BGE-reranker (Xiao et al., 2024).

All retrieval methods return the top-k sentences (k=7 for retrieval, k=5 for reranking). Notably, SBERT, BioLinkBERT, and MedCPT produce identical scores (Table 2); with only ∼21 sentences per note on average, these dense retrievers return largely overlapping top-7 sets despite different embedding spaces.

4.1.2. LLM Prompting

We design four prompt variants that differ in their framing of the selection task:
Zero-shot: A minimal prompt asking the model to identify relevant sentences and output a JSON array.
Inclusive: Instructs the model to identify *ALL* relevant sentences, emphasizing "it's better to include a marginally relevant sentence than to miss important evidence." This framing optimizes for recall.
Balanced: Uses a three-tier categorization (essential, supporting, background) to guide selection granularity, balancing precision and recall.
Stepwise (CoT): Decomposes the task into four explicit steps (identify key concepts, locate relevant sentences, find supporting context, and compile the final list), eliciting chain-of-thought reasoning (Wei et al., 2022).

4.1.3. Novel Methods

SCEV (Self-Consistency Evidence Verification): Inspired by self-consistency (Wang et al., 2023), we generate 5 evidence selections using the balanced prompt at varying temperatures (0.3 to 0.9), then apply majority voting with a 40% threshold. Sentences selected by at least 2 of 5 samples are retained.
CEA (Counterfactual Evidence Attribution): We first obtain a broad candidate set using inclusive prompting. For each candidate sentence, we test its importance by removing it from the note and asking whether the question can still be fully answered. Sentences whose removal renders the question unanswerable are classified as essential.
Question Decomposition: We decompose the patient question into sub-questions and independently identify evidence for each, then take the union.
Claim Verification: We extract claims from the question, then verify each claim against individual sentences to find supporting evidence.
Multi-Perspective: We prompt from three perspectives (clinical accuracy, completeness, patient relevance) and combine selections via union.

4.1.4. Hybrid and Ensemble Approaches

BioLinkBERT + LLM: Pre-filter to top-15 with BioLinkBERT, then use LLM to refine.
SBERT + LLM: Pre-filter to top-15 with SBERT, then use LLM to refine from this candidate set.
Ensemble Voting: Run three prompts (zero-shot, inclusive, balanced) and retain sentences selected by at least one method.
Iterative Refinement: Generate initial selection, self-critique for missed evidence and false positives, then produce a revised selection.

4.2. Subtask 3: Answer Generation

We evaluate 13 methods spanning direct prompting, multi-stage pipelines, a Constitutional AI approach, and DSPy-optimized modules.

4.2.1. Direct Prompting Methods

Constrained: The required baseline that instructs the model to answer using *only* information from the clinical notes, with a 75-word limit and five explicit constraints (no external knowledge, professional terminology, acknowledge limitations).
Zero-shot: Minimal prompt with just the question and notes.

Balanced: Adds the clinician's interpretation and structured instructions for balanced coverage.

4.2.2. Multi-Stage Methods

CoT-Verified: Two-stage pipeline: (1) chain-of-thought generation with explicit identify, locate, synthesize, verify steps, followed by (2) self-verification against the notes.
Evidence-First: Leverages Subtask 2 output by first extracting evidence sentences, then generating the answer from only the identified evidence.
Iterative Refinement: Three-stage generate, critique, revise pipeline using separate prompts for each stage.
Multi-Perspective: Elicits responses from three simulated experts (conservative, comprehensive, patient-centered), then synthesizes a final answer.
Claim-Verified: Extracts individual claims with evidence citations, then synthesizes verified claims into a coherent answer.
Decompose-Synthesize: Decomposes the question into 2 to 4 sub-questions, answers each independently from the notes, then synthesizes a unified response.

4.2.3. Constitutional AI Pipeline

Our approach adapts the principle-guided critique-and-revise mechanism from Constitutional AI (Bai et al., 2022) as an inference-time prompting strategy for clinical faithfulness. The pipeline has three stages (Figure 2):
Stage 1 (Generate): Produce an initial answer using the constrained prompt.
Stage 2 (Critique): Evaluate the draft against five clinical principles:

1. All claims must be directly traceable to the clinical notes
2. Do not speculate or make inferences beyond documentation
3. Acknowledge when information is incomplete
4. Use exact medical terminology from the notes
5. Maintain temporal accuracy (do not confuse past/present)

The model evaluates each principle as PASS or FAIL with explanations.
Stage 3 (Revise): Generate a revised answer that addresses all principle violations while maintaining the 75-word constraint.

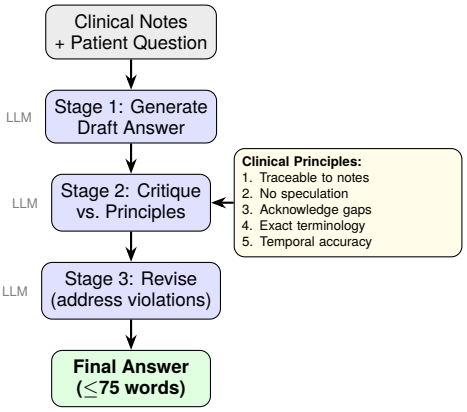

Figure 2: Constitutional AI pipeline for clinical answer generation. Each stage uses a separate LLM call. The five principles enforce clinical faithfulness during the critique stage.

4.2.4. DSPy Optimization

Following the success of DSPy at ArchEHR-QA 2025 (Bogireddy et al., 2025), we implement three DSPy modules optimized with MIPROv2 (Opsahl-Ong et al., 2024):
DirectQA: Single-stage ChainOfThought answer generation.
TwoStageQA: First extracts evidence sentence IDs, then generates from the extracted evidence.
SelfConsistencyQA: Generates 5 candidate answers and selects the most consistent one.

All modules are optimized on the 20-example dev set using MIPROv2 with the "light" auto-configuration. The optimization metric combines ROUGE-Lsum similarity between the generated and gold answers with a penalty for exceeding the 75-word limit.

5. Experimental Setup

5.1. Models

We evaluate across models from two providers:
OpenAI (API): GPT-5, GPT-4o, and GPT-5.2 (an updated GPT-5 variant with improved instruction following) (OpenAI, 2025).
Anthropic (AWS Bedrock): Claude Opus 4.5 and Claude Haiku 4.5 (Anthropic, 2025).

Not all models are evaluated on all methods due to API cost constraints; Tables 3 and 4 report the most informative subset. Opus was evaluated only on zero-shot for Subtask 2.

5.2. Hyperparameters

For Subtask 2, LLM methods use temperature 0.0 for deterministic output. Retrieval methods use top-k=7 (retrieval) or k=5 (reranking). SCEV uses 5 samples at temperatures 0.3 to 0.9 with a 40%

voting threshold. For Subtask 3, generation uses temperature 0.3 with max_tokens = 500, and all outputs are truncated to 75 words.

5.3. Submission

For Subtask 2, we submit the inclusive prompt with GPT-5, prioritizing recall. We selected lenient F1 over strict F1 as the evaluation criterion because it assigns partial credit for supplementary evidence, better reflecting the downstream utility of evidence retrieval: for answer generation (Subtask 3), including supplementary context sentences is preferable to missing them entirely. Although other prompts achieve higher lenient F1 (e.g., stepwise: 69.0), the inclusive prompt's near-complete recall (90.1% vs. 64.5% for balanced) minimizes the risk of missing evidence that benefits Subtask 3. For Subtask 3, we submit the constitutional method with Claude Opus 4.5 based on the best dev overall score.

6. Results and Analysis

6.1. Subtask 2: Evidence Identification

6.1.1. Dev Set Results

Table 2 shows the dev set results comparing non-LLM baselines and LLM methods. LLM prompting methods substantially outperform retrieval baselines. The best retrieval method (SBERT/BioLinkBERT, 46.7 strict F1) is surpassed by the best LLM prompting (stepwise GPT-5, 65.9 strict F1), a gap of 19.2 points. Even the simplest zero-shot prompt (64.2) outperforms all retrieval baselines.

6.1.2. Cross-Model Comparison

Table 3 shows how prompt strategies interact with different models. The stepwise (CoT) prompt shows the most dramatic model-dependency: it achieves 65.9 F1 with GPT-5 but only 16.7 with Claude Haiku 4.5, a catastrophic 49-point drop. This is because Haiku, a smaller model, struggles with the multi-step reasoning format, often failing to produce valid JSON output after the reasoning chain.

6.1.3. Test Results

Our test submission uses inclusive prompting with GPT-5, achieving **58.0 lenient micro F1** (rank 12). As discussed in Section 5.3, we prioritized recall (90.1% on dev) over F1 to maximize evidence coverage for downstream answer generation, accepting the trade-off of lower precision (45.4%).

Method	Model	S-F1	S-P	S-R
Non-LLM Retrieval Baselines				
BM25	–	39.1	36.4	42.1
TF-IDF	–	41.4	38.6	44.6
SBERT	–	46.7	43.6	50.4
BioLinkBERT	–	46.7	43.6	50.4
MedCPT	–	46.7	43.6	50.4
Clin-Longformer	–	40.6	37.9	43.8
MiniLM+CE	–	38.0	42.0	34.7
BGE+Rerank	–	43.4	48.0	39.7
LLM Prompting Methods (GPT-5)				
Zero-shot	GPT-5	64.2	63.9	64.5
Inclusive	GPT-5	60.4	45.4	**90.1**
Balanced	GPT-5	65.8	**67.2**	64.5
Stepwise (CoT)	GPT-5	**65.9**	59.2	74.4
Novel Methods (GPT-5)				
SCEV	GPT-5	64.4	60.0	69.4
CEA	GPT-5	46.5	46.0	47.1
Decomposition	GPT-5	42.3	39.6	45.5
Claim Verify	GPT-5	50.9	57.9	45.5
Multi-persp.	GPT-5	60.6	54.2	68.6
Hybrid / Ensemble (GPT-5)				
BioLink+LLM	GPT-5	55.5	54.8	56.2
SBERT+LLM	GPT-5	53.8	69.7	43.8
Ensemble	GPT-5	58.1	41.5	96.7
Iterative	GPT-5	58.4	48.1	74.4

Table 2: Subtask 2 dev results (strict micro scores). S-F1 = strict micro F1, S-P = precision, S-R = recall. Best per column in **bold**.

Prompt	GPT-5	5.2	4o	Opus	Haiku
Zero-shot	**64.2**	53.2	49.8	55.5	52.6
Inclusive	60.4	58.5	59.5	–	60.8
Balanced	65.8	59.9	56.4	–	57.9
Stepwise	65.9	59.4	53.2	–	16.7
SCEV	64.4	60.2	58.9	–	59.3

Table 3: Subtask 2 cross-model comparison (strict micro F1). Opus = Claude Opus 4.5, Haiku = Claude Haiku 4.5. GPT-5 consistently leads.

6.2. Subtask 3: Answer Generation

6.2.1. Dev Set Results

Table 4 shows the dev set results for Subtask 3. The Constitutional AI method with Claude Opus 4.5 achieves the highest overall score (31.4), though the margin over the constrained baseline (30.7) is modest given the small dev set size (20 cases), and the differences may not be statistically significant. The improvement is primarily driven by BLEU (+1.18) and ROUGE-Lsum (+1.67). Notably, BERTScore is remarkably stable across methods (34.6 to 40.3), while BLEU varies from 1.77 to 6.29, suggesting that BERTScore is less discriminative for this task and that overall score differences are driven primarily by lexical overlap metrics. Note that

AlignScore and MEDCON were not computed during dev evaluation (both zero for all methods), so the dev overall score averages only BLEU, ROUGE-Lsum, SARI, and BERTScore.

6.2.2. Test Results

Our test submission uses the constitutional method with Claude Opus 4.5, achieving an overall score of **31.8** (rank 10).

6.3. Analysis

6.3.1. LLMs vs. Retrieval for Evidence Identification

The most striking finding from Subtask 2 is the large gap between retrieval baselines and LLM prompting. The best retrieval method achieves 46.7 strict F1, while the best LLM prompt (stepwise GPT-5) reaches 65.9, a gap of 19.2 points. This suggests that LLMs can effectively perform the sentence-level relevance assessment required for clinical evidence identification, leveraging their understanding of clinical context in ways that embedding similarity alone cannot capture.

6.3.2. Prompt Framing and Precision/Recall

Our four prompt variants reveal a clear precision/recall spectrum controlled entirely by framing:

- **Inclusive** (R=90.1, P=45.4): Maximizes recall by instructing "be INCLUSIVE."
- **Stepwise/CoT** (R=74.4, P=59.2): Structured reasoning improves precision while maintaining high recall.
- **Zero-shot** (R=64.5, P=63.9): Balanced default behavior.
- **Balanced** (R=64.5, P=67.2): Three-tier categorization yields the highest precision.

This finding has practical implications: practitioners can tune the precision/recall trade-off purely through prompt design, without any model retraining. In clinical settings, inclusive framing is preferred when missed evidence carries high risk (e.g., safety-critical questions about medications or procedures), while balanced framing suits scenarios where false positives are costly (e.g., automated summarization where irrelevant content degrades output quality).

6.3.3. Chain-of-Thought is Model-Dependent

The stepwise (CoT) prompt exhibits dramatic model-dependency. With GPT-5, it achieves the highest F1 (65.9) of any single prompt. With Claude Haiku 4.5, it produces only 16.7 F1, a catastrophic failure. Analysis of Haiku's outputs reveals that the smaller model often fails to complete the reasoning chain, producing malformed JSON or getting stuck in verbose reasoning without reaching the final answer format. Intermediate models (GPT-5.2, GPT-4o) show modest benefits from CoT, suggesting that the effectiveness of structured reasoning scales with model capability. Practitioners should be cautious with CoT prompting for smaller models; in our experiments, Claude Haiku catastrophically fails at structured reasoning, while GPT-4o handles it adequately but with degraded performance relative to simpler prompts.

To illustrate, consider Case 4 (a patient asking about cardiac catheterization for heart failure; 7 essential and 2 supplementary gold evidence sentences). The inclusive prompt with GPT-5 retrieves 12 sentences, capturing 8 of 9 gold sentences (missing only sentence 10). In contrast, the stepwise prompt returns only sentence 5: the multi-step reasoning caused the model to over-filter, retaining just one of nine relevant sentences. While stepwise achieves the best overall F1 across the dev set, its performance is highly variable per case; Case 4 represents a worst-case failure mode where chain-of-thought over-filtered. See Appendix B for the full comparison.

These results suggest practical conditions for CoT reliability: structured reasoning prompts should be reserved for high-capability models (here, GPT-5 class or above) that can maintain coherent multi-step reasoning while adhering to output format constraints. For smaller or less capable models, the added complexity of CoT increases the risk of format violations (e.g., malformed JSON) and reasoning failures that degrade performance below simpler prompts. Practitioners should validate CoT effectiveness on their target model before deployment rather than assuming transferability across model scales.

6.3.4. Principle-Guided Critique Improves Answer Quality

The Constitutional AI pipeline achieves the best overall score on both dev (31.4) and test (31.8) for Subtask 3, though the dev margin (31.4 vs. 30.7) is small and may not be significant given only 20 evaluation cases. Comparing constitutional to constrained (same model, Claude Opus 4.5), the critique-and-revise cycle improves BLEU by +1.18 (5.11 $\rightarrow$ 6.29) and ROUGE-Lsum by +1.67 (19.8 $\rightarrow$ 21.5), while SARI remains stable (58.5).

Analysis of principle violations across the 20 dev cases reveals that Principle 1 (traceability to notes) is the most frequently violated, failing in 9 of 20 drafts (45%), followed by Principle 2 (no speculation) which often co-occurs with Principle 1 viola-

Method	Model	Ovr.	BLEU	R-Lsum	SARI	BERTSc.
Constitutional	Opus 4.5	**31.4**	**6.29**	**21.5**	**58.5**	39.2
Constrained	Opus 4.5	30.7	5.11	19.8	58.5	39.3
DSPy-SC	GPT-4o	30.6	5.23	21.4	56.9	38.8
DSPy-Direct	GPT-4o	30.5	4.62	20.9	56.2	**40.3**
DSPy-SC	GPT-5.2	30.4	4.42	21.5	56.3	39.4
Constrained†	Opus 4.5	30.4	4.55	19.8	58.3	38.9
DSPy-Direct	GPT-5.2	29.9	4.07	20.7	57.1	37.9
DSPy-SC	Opus 4.5	29.0	2.39	19.0	55.6	39.1
Iterative	Opus 4.5	27.9	2.53	18.2	55.3	35.5
DSPy-2Stage	GPT-4o	26.7	1.77	18.7	51.9	34.6

Table 4: Answer generation dev results sorted by overall score. R-Lsum = ROUGE-Lsum, BERTSc. = BERTScore. Overall averages BLEU, R-Lsum, SARI, and BERTScore (AlignScore and MEDCON were zero for all methods during dev evaluation). †Earlier run with different random seed. DSPy-SC = SelfConsistency.

tions. Principles 3 to 5 are rarely triggered. Common error patterns include adding unsupported severity qualifiers ("severe," "significant"), speculating about causal mechanisms not documented in the notes, and making inferences beyond what is explicitly stated. Table 5 illustrates how the critique stage identifies these violations and the revision replaces speculation with documented facts.

However, in cases where the draft already passes all principles (11 of 20 dev cases), the critique-and-revise cycle adds computational cost without meaningful improvement. Additionally, the 75-word constraint means revisions sometimes drop relevant information to accommodate corrections, trading completeness for faithfulness. We discuss these cost trade-offs further in the Limitations section.

6.3.5. DSPy Optimization Performance

DSPy-optimized modules with MIPROv2 produce competitive results (30.5 to 30.6 overall) but do not meaningfully improve over the constrained baseline (30.7 with Claude Opus 4.5). Given the small dev set (20 examples), the differences among DSPy, constrained, and constitutional methods are within the noise margin and likely not statistically significant. This suggests that automatic prompt optimization with MIPROv2 on a 20-example dev set provides insufficient signal to outperform a well-designed manual prompt for this task. The SelfConsistencyQA module slightly outperforms DirectQA (30.6 vs. 30.5 with GPT-4o), consistent with the self-consistency literature (Wang et al., 2023). The TwoStageQA module underperforms (26.7), suggesting that the two-stage pipeline introduces additional complexity without corresponding gains at this scale.

7. Discussion

7.1. Practical Implications

Our systematic comparison provides practical guidance for clinical QA system design. For evidence identification, simple LLM prompting with clear framing outperforms complex retrieval pipelines, hybrid systems, and novel multi-step approaches. The inclusive prompt's high recall makes it suitable when downstream tasks (e.g., answer generation) benefit from comprehensive evidence, while balanced prompting is preferable for precision-critical applications.

For answer generation, the Constitutional AI approach demonstrates that domain-specific self-critique can improve answer quality at modest computational cost (3 LLM calls vs. 1). The five clinical principles provide an interpretable framework that could be extended with additional domain expertise.

To guide practitioners in selecting a prompt strategy: inclusive framing (high recall) is recommended when downstream tasks depend on comprehensive evidence coverage, for example, when generating patient-facing answers where missing a relevant clinical detail could lead to an incomplete or misleading response. Balanced framing (high precision) is preferable in precision-critical settings such as clinical decision support, where false positives could waste clinician attention or introduce noise into automated workflows. The zero-shot prompt provides a reasonable default when no strong prior exists about the precision/recall requirements of the application.

7.2. Dev/Test Gap

The constitutional method scores 31.4 on dev and 31.8 on test, though these are not directly comparable: the dev overall averages 4 metrics (AlignScore and MEDCON were not computed), while the test

Case 2: *Why was father given Lasix?*	**Case 4:** *Why cardiac catheterization?*	**Case 7:** *Why was Coumadin stopped?*
Draft (red = unsupported claims)		
"...acute diastolic heart failure with significant volume overload...and severe lower extremity edema. [...] His oxygen was likely reduced as diuresis improved his respiratory status."	"...revealed elevated filling pressures...and reduced cardiac output (Fick cardiac index of 1.72), confirming severe low-output heart failure. The milrinone trial during catheterization was successful..."	"...Coumadin was discontinued following your subarachnoid hemorrhage and IVC filter placement. [...] The discontinuation is directly related to your brain hemorrhage, as anticoagulation poses bleeding risks in this context."
Critique (principle violations identified)		
P1 FAIL: "severe" not in notes; causal relationship "likely reduced as diuresis improved" not stated	P1 FAIL: notes say "low-output" without the qualifier "severe"; P2 FAIL: "confirming" implies diagnostic certainty not stated	P1 FAIL: reason for discontinuation not explicitly stated; P2 FAIL: "anticoagulation poses bleeding risks" is external medical reasoning
Revision (teal = corrected/added from notes)		
"...acute diastolic heart failure with volume overload...and lower extremity edema. The Lasix was administered intravenously, including via continuous drip, then transitioned to oral torsemide to remove excess fluid."	"...showed elevated filling pressures (mean PCW of 30) and reduced cardiac output (Fick C.I. of 1.72), consistent with low output heart failure. The milrinone trial was successful, improving your Fick C.I. to 2.79."	"...Coumadin was discontinued following your IVC filter placement. [...] The clinical notes do not explicitly state the reason for discontinuation beyond documenting it occurred after filter placement."

Table 5: Constitutional AI examples across three cases. Red text in drafts marks unsupported claims flagged by the critique. Teal text in revisions highlights corrections: added documented details, softened language, or explicit acknowledgment of information gaps.

overall averages all 6 metrics. For Subtask 2, the dev/test gap is larger: the inclusive prompt achieves 64.3 lenient micro F1 on dev (60.4 strict) but 58.0 lenient on test. This may reflect the test set's greater diversity (47 vs. 20 cases) or distributional shift in question types and clinical note complexity.

8. Conclusion

We presented GigitAI's systems for ArchEHR-QA 2026 Subtasks 2 and 3, featuring a systematic comparison of 21 evidence identification methods and 13 answer generation methods across 5 language models. Our key findings are that LLM prompting outperforms traditional retrieval by a wide margin (19 F1 points) for clinical evidence identification, prompt framing provides fine-grained control over precision/recall trade-offs, chain-of-thought effectiveness is strongly model-dependent, and Constitutional AI with clinical principles can improve answer generation quality by catching unsupported claims in LLM drafts. Our best submissions achieve 58.0 F1 (rank 12) for evidence identification and 31.8 overall (rank 10) for answer generation.

9. Limitations

Our study has several limitations. First, the small dev set (20 cases) limits our ability to draw statistically robust conclusions; even the same method with different random seeds varies by 0.3 points (Table 4), illustrating the measurement noise at this scale. Second, we rely entirely on prompting and do not explore fine-tuning, which could improve performance, particularly for smaller models where CoT prompting fails. Third, our multi-stage pipelines (Constitutional AI, CoT-Verified, Iterative Refinement) require 2 to 3 LLM calls per answer, multiplying both latency and cost relative to single-call methods. For the Constitutional AI pipeline specifically, the critique-and-revise cycle adds no value in 55% of cases (11/20 dev cases pass all principles on the first draft), making the additional cost wasteful for already-faithful outputs. A confidence-based gating mechanism could skip the critique stage when the draft is likely faithful, reducing average cost. Fourth, we rely entirely on proprietary LLM APIs (OpenAI, Anthropic), which limits reproducibility and raises concerns about cost, data privacy, and access continuity. Evaluating open-weight models (e.g., Llama (Grattafiori et al., 2024), Mistral (Jiang et al., 2023)) would strengthen the generalizability of our findings. Fifth, the 75-word constraint for Subtask 3 makes it difficult to produce comprehensive answers for complex clinical questions. Finally, our clinical principles are manually designed; automatic principle discovery could further improve the approach.

10. Bibliographical References

Anthropic. 2025. Claude opus 4.5 and claude haiku 4.5 system cards. Technical report, Anthropic.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. ArXiv:2212.08073 [cs].

Sai Prasanna Teja Reddy Bogireddy, Abrar Majeedi, Viswanath Gajjala, Zhuoyan Xu, Siddhant Rai, and Vaishnav Potlapalli. 2025. Neural at ArchEHR-QA 2025: Agentic prompt optimization for evidence-grounded clinical question answering. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 104–109, Vienna, Austria. Association for Computational Linguistics.

Ary Goldberger, Luís Amaral, Leon Glass, Jeffrey Hausdorff, Plamen Ivanov, Roger Mark, Joseph Mietus, George Moody, Chung-Kang Peng, and H. Stanley. 2000. Physiobank, physiotoolkit, and physionet : Components of a new research resource for complex physiologic signals. *Circulation*, 101:E215–20.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. MIMIC-IV. *PhysioNet*. Version 3.1.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines. In *The Twelfth International Conference on Learning Representations*.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Hyun Kim, Seunghyun Won, and Edward Choi. 2024. EHRNoteQA: A Patient-Specific Question Answering Benchmark for Evaluating Large Language Models in Clinical Settings. *PhysioNet*. Version 1.0.0.

Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *CoRR*, abs/2201.11838.

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

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

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas

King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine.

OpenAI. 2025. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*.

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.

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, Brussels, Belgium. 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*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 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.

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.

Arun Thirunavukarasu, Darren Ting, Kabilan Elangovan, Laura Gutierrez Sinisterra, Ting Tan, and Daniel Ting. 2023. Large language models in medicine. *Nature Medicine*, 29.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

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.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Wen-wai Yim, Yujuan Velvin 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.

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. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A. Prompt Templates

This appendix lists the key prompt templates used in our systems. Placeholders are shown in `{braces}`.

A.1. Evidence Identification (Subtask 2)

Inclusive Prompt:

> *You are a clinical expert identifying relevant evidence from medical records.*
>
> *Patient Question:* `{patient_question}`
> *Clinician's Interpretation:* `{clinician_question}`
>
> *Clinical Note Sentences:* `{numbered_sentences}`
>
> *Your task: Identify ALL sentences that contain information relevant to answering the question.*
>
> *IMPORTANT: Be INCLUSIVE. It's better to include a marginally relevant sentence than to miss important evidence.*
>
> *Output a JSON array of ALL relevant sentence numbers.*

Balanced Prompt:

> *Identify sentences that contain information needed to answer the question.*
>
> *Selection criteria: 1. ESSENTIAL: Sentences that directly answer part of the question (MUST include) 2. SUPPORTING: Sentences that provide important context or values (SHOULD include) 3. BACKGROUND: Sentences that help understand the clinical situation (MAY include)*

Stepwise (CoT) Prompt:

> *Let's identify evidence step by step: Step 1: What are the key clinical concepts in the question? Step 2: Which sentences contain information about these concepts? Step 3: Which sentences provide supporting context or values? Step 4: Compile the final list of relevant sentences.*
>
> *Think through each step, then output ONLY a JSON array of sentence numbers.*

A.2. Answer Generation (Subtask 3)

Constrained Prompt:

> *You are a clinical expert answering a patient's question about their hospitalization.*
>
> *Patient Question:* `{patient_question}`
> *Clinician's Interpretation:* `{clinician_question}`
> *Clinical Notes:* `{note_excerpt}`
>
> *Instructions: 1. Answer ONLY using information from the clinical notes above 2. Use professional medical terminology (not overly simplified) 3. If information is incomplete, acknowledge limitations 4. Maximum 75 words, approximately 5 sentences 5. Do NOT speculate or add external medical knowledge*

Constitutional AI, Critique Prompt:

> *Clinical Principles: 1. All claims must be directly traceable to the clinical notes 2. Do not speculate or make inferences beyond documentation 3. Acknowledge when information is incomplete 4. Use exact medical terminology from the notes 5. Maintain temporal accuracy (don't confuse past/present)*
>
> *Review this answer against the clinical principles above.*
>
> *Clinical Notes:* `{note_excerpt}`
> *Draft Answer:* `{draft}`
>
> *For each principle, evaluate compliance (PASS/FAIL) and explain:*

Constitutional AI, Revision Prompt:

> *Revise this answer to comply with all clinical principles.*
>
> *Original Answer:* `{draft}`
> *Principle Evaluation:* `{critique}`
> *Clinical Notes (for reference):* `{note_excerpt}`
>
> *Revised answer (must comply with ALL principles, under 75 words):*

B. Evidence Identification Example

Table 6 shows a detailed comparison of evidence identification methods on Case 4, where the patient asks *"Why was cardiac catheterization recommended?"*. Gold evidence includes 7 essential and 2 supplementary sentences across the discharge summary. We compare four methods: BM25 (sparse retrieval), SBERT (dense retrieval), inclusive prompt (GPT-5), and stepwise/CoT prompt (GPT-5).

The inclusive prompt captures 8 of 9 gold sentences (missing only sentence 10), while stepwise/CoT retrieves only 1 sentence, demonstrating the over-filtering failure mode discussed in Section 6.3. BM25 captures 4 of 9 gold sentences and SBERT captures 5 of 9, but they select different subsets: BM25 favors lexically similar sentences (e.g., "milrinone"), while SBERT favors semantically similar ones (e.g., "heart failure").

ID	Gold	Sentence (truncated)	BM25	SBERT	Inclusive	Stepwise
2		Cardiology service: abdominal pain, nausea, vomiting felt to be secondary to congestive hepatopathy	✓		✓	
4		He was aggressively diuresed with a net 10 liters negative since admission	✓			
5	**Ess.**	He underwent RHC for milrinone trial, which proved to be successful	×	×	✓	✓
6		His mean PCW went from 30 to 22, and his Fick C.I. went from 1.72 to 2.79			✓	
7	**Supp.**	Brief Hospital Course: 48M with idiopathic dilated cardiomyopathy (EF 25%…)	×	✓	✓	×
8		Acute-on-chronic systolic heart failure		✓	✓	
9	**Supp.**	Underlying exacerbation of chronic heart failure, known idiopathic cardiomyopathy…	×	×	✓	×
10	**Ess.**	Last echo showed LVEF = 25%	×	×	×	×
11	**Ess.**	Admitted in the setting of low output heart failure, leading to increased intra-abdominal pressures…	✓	×	✓	×
13	**Ess.**	Cardiac output and wedge pressure significantly improved after milrinone infusion	✓	✓	✓	×
14		In the ICU maintained on milrinone at 0.5 mcg/hr, transferred to floor	✓			
16		Heart failure specialists had honest discussions about long-term prognosis		✓	✓	
18	**Ess.**	You were admitted to the hospital with worsening heart failure	✓	✓	✓	×
19	**Ess.**	You had a cardiac catheterization that showed you would benefit from milrinone	✓	✓	✓	×
20	**Ess.**	You were started on a milrinone drip, with improvement in your heart's pump function	×	✓	✓	×
Gold sentences found / total gold (9)			4/9	5/9	**8/9**	1/9
False positives			3	2	4	0

Table 6: Evidence identification comparison on Case 4 (*"Why was cardiac catheterization recommended?"*). Shaded rows are gold evidence (Ess. = essential, Supp. = supplementary). ✓ = correctly identified gold sentence; × = missed gold sentence; ✓ = selected but not gold (false positive). Only sentences selected by at least one method are shown.