WisPerMed at ArchEHR-QA 2026: Retrieval-Augmented Prompting for Grounded EHR Question Answering

Jan-Henning Büns[1], Tabea M. G. Pakull[2,1], Hendrik Damm[1,3], Bohao Chu[6], Christoph M. Friedrich[1,7], Felix Nensa[4,5], Elisabeth Livingstone[2], Peter A. Horn[2], Norbert Fuhr[6]

[1]Department of Computer Science, University of Applied Sciences and Arts Dortmund
[2]Institute for Transfusion Medicine, University Hospital Essen
[3]Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen
[4]Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen
[5]Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen
[6]Department of Computer Science, University of Duisburg-Essen
[7]Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen
Correspondence: jan-henning.buens@fh-dortmund.de

Abstract

ArchEHR-QA is a grounded question-answering (QA) task for electronic health records (EHRs) comprising four subtasks: (1) question rewriting, (2) evidence identification, (3) grounded answer generation, and (4) answer-evidence alignment. In this work, we present a modular pipeline centered on retrieval-augmented generation (RAG). For Subtask 1, RAG few-shot prompting outperformed both PEFT and prompt-only baselines on the development set; however, Claude few-shot proved substantially more robust on the test set, ranking 6th out of 6 participating teams (score: 26.94). For Subtask 2, a union ensemble of open-weight LLMs (GPT-OSS-120B and Qwen3-30B-A3B) achieved a 56.7 micro-F1, rivaling the proprietary Claude Opus 4.6 while demonstrating higher recall (53.6). For Subtask 3, our RAG few-shot approach using Claude Opus 4.5 achieved the 1st place out of 13 participating teams (score: 36.33). Finally, for Subtask 4, a zero-shot Claude Opus 4.6 configuration ranked 2nd out of 16 participating teams (score: 81.3).

Keywords: Electronic Health Records, Question Answering, Retrieval-Augmented Generation, Large Language Models, Clinical NLP

1. Introduction

Patient portals and electronic messaging systems are fundamentally changing how patients interact with their own health information. As these platforms become widespread, patients increasingly access clinical documentation—discharge summaries, progress notes, laboratory results—and pose questions to their care teams through asynchronous messaging (Johnson et al., 2016). This shift creates a growing demand for automated systems that can assist clinicians in responding to patient inquiries accurately and efficiently. However, building such systems is far from straightforward: patient questions are often verbose, colloquial, and underspecified, while the clinical notes that contain the answers are long, heterogeneous, and densely packed with medical terminology across multiple concurrent problems.

A key requirement for any question-answering (QA) system operating in this clinical setting is *grounding*—the ability to not only generate fluent and medically appropriate responses, but to explicitly link each claim in the response to supporting evidence in the patient's electronic health record (EHR). Without grounding, even factually correct answers lack the verifiability that clinicians need to trust and adopt automated assistance, and hallucinated content poses a direct risk of patient harm. This requirement distinguishes clinical QA from open-domain question answering, where fluency and factual accuracy are often sufficient.

The ArchEHR-QA shared task (Soni and Demner-Fushman, 2026b) formalizes these challenges into four complementary subtasks that together constitute a complete grounded QA pipeline: (1) *question interpretation*, which rewrites a verbose patient question into a concise clinician query; (2) *evidence identification*, which selects the minimal set of EHR sentences needed to answer the question; (3) *answer generation*, which produces a grounded response constrained by the clinical evidence; and (4) *answer–evidence alignment*, which maps each answer sentence to its supporting evidence at the sentence level. This decomposition reflects the real-world workflow of clinicians who must first understand the patient's intent, locate the relevant information, formulate a response, and ensure that every statement is supported.

In this work, we present a modular pipeline for all four subtasks, built around large language models (LLMs) with retrieval-augmented generation (RAG). Our system design is guided by three practical constraints that are central to clinical NLP deployment. First, *format fidelity*: each subtask imposes strict

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 455–468
12 May 2026.

output constraints (e.g., at most 15 words for question rewriting, at most 75 words for answer generation) that must be enforced reliably. Second, *faithfulness*: generated answers must be grounded in the provided clinical notes without introducing unsupported claims. Third, *reproducibility and data privacy*: clinical data is subject to stringent governance requirements, motivating a systematic comparison between locally deployable open-weight LLMs and proprietary API-based models to quantify the accuracy–privacy trade-off. To support independent development, error analysis, and auditability, each subtask is addressed by a dedicated module that can be evaluated and improved in isolation.

Contributions. Our main contributions are:

- A **RAG in-context learning strategy** that dynamically retrieves semantically similar exemplars and enforces task-specific output constraints for question rewriting and answer generation, ranking 1st out of 13 teams on answer generation.
- A **privacy-aware open-LLM ensemble** for evidence identification, combining two open-weight LLMs via a union strategy that achieves competitive performance with a proprietary LLM (56.7 vs. 58.8 micro-F1) while providing higher recall and preserving full data sovereignty.
- Evidence that **patient question context substantially improves answer–evidence alignment**, with a zero-shot prompting approach ranking 2nd out of 16 teams (micro-F1: 81.3).
- A **systematic analysis across all four subtasks**, comparing prompting, RAG, fine-tuning, and preference optimization paradigms, with a detailed examination of how distribution shift degrades retrieval-based methods relative to stronger LLM backbones.

2. Related Work

Grounded QA from clinical text. Evidence-grounded (QA) has been extensively studied in open-domain settings, but grounding answers in EHR notes introduces additional challenges: long and heterogeneous narratives, multiple concurrent problems, and high consequences of hallucinated content. The ArchEHR-QA shared task formalizes these challenges with explicit evidence annotations and answer–evidence alignments (Soni and Demner-Fushman, 2026b). The dataset is derived from de-identified clinical documentation in the style of MIMIC critical care records, and thus inherits the linguistic variability and documentation conventions typical of real-world EHR notes (Johnson et al., 2016).

Consumer health question rewriting and summarization. Subtask 1 is closely related to consumer health question summarization, where verbose patient queries are condensed while preserving intent and clinical content. MeQSum provides expert summaries for such questions and is widely used for training and evaluating medical question summarization systems (Ben Abacha and Demner-Fushman, 2019). Unlike generic rewriting, ArchEHR-QA requires the output to resemble a clinician query to an EHR system, without introducing unsupported content.

RAG prompting. RAG (Lewis et al., 2020) and retrieval of in-context demonstrations (e.g., nearest-neighbor prompting) can improve few-shot performance by providing style- and content-aligned exemplars at inference time (Shi et al., 2022). In parallel, parameter-efficient fine-tuning methods such as LoRA reduce the cost of adapting LLMs to domain-specific tasks (Hu et al., 2022), while preference optimization (e.g., DPO) can shift LLM behavior toward desired outputs without explicit reward modeling (Rafailov et al., 2023).

3. Data

Each case in the ArchEHR-QA dataset (Soni and Demner-Fushman, 2026a) contains a patient-authored question, a clinician-interpreted question (reference for Subtask 1), a clinical note excerpt segmented into numbered sentences with sentence-level relevance annotations, a clinician-authored answer (reference for Subtask 3), and answer–evidence alignments (reference for Subtask 4) (Soni and Demner-Fushman, 2026b). The dataset is distributed via PhysioNet and is derived from de-identified EHR documentation (MIMIC-style) (Johnson et al., 2016).

4. System Overview

Our system follows a modular pipeline aligned with the subtasks of ArchEHR-QA (Organizers, 2026).

Subtask 1: Question Interpretation. Given a patient-authored question, the system must produce a single clinician-interpreted question that captures the core information need in at most 15 space-separated words, ends with a question mark, and does not introduce facts absent from the input. We address this with RAG few-shot prompting, dynamically selecting style-matched exemplars from the training set to guide format and clinical register.

Subtask 2: Evidence Identification. Given a patient question, a clinician question (gold or system-generated), and a numbered clinical note excerpt,

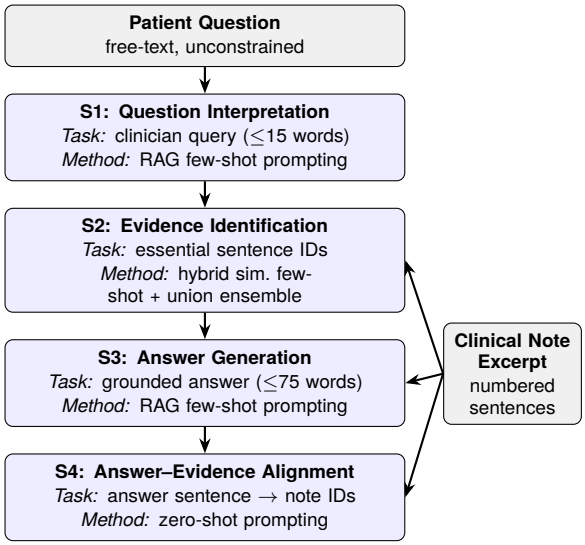

Figure 1: Modular pipeline for ArchEHR-QA. Each box states the task output constraint and the method used.

the system must identify the minimal set of sentence IDs that sufficiently support answering the question. We approach this as a direct selection task using hybrid similarity-based few-shot in-context learning, and combine predictions from two open-weight LLMs via a union ensemble that exploits their complementary precision–recall profiles.

Subtask 3: Answer Generation. Given a patient question, a clinician question, and a clinical note excerpt, the system must generate a grounded answer in professional register, limited to at most 75 words and avoiding speculation where evidence is partial. We generate answers conditioned on the full note excerpt with evidence emphasis, using RAG few-shot prompting to supply style- and content-aligned demonstrations at inference time.

Subtask 4: Answer–Evidence Alignment. Given a patient question, a numbered clinical note excerpt, and a numbered answer text, the system must map each answer sentence to the set of note sentences that support it, producing many-to-many alignments. We adopt a zero-shot prompting strategy, instructing the LLM to identify supporting evidence sentences for each answer sentence individually.

5. Selection of LLMs

We primarily use open-weight instruction-tuned LLMs for local inference and reproducibility: Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) and OpenAI's open-weight gpt-oss-120b (OpenAI, 2025), served via vLLM (Kwon et al., 2023). For all subtasks comparisons (patient question only), we also report closed-LLM baselines using Claude Opus 4.5 (Anthropic, 2025) and Claude Opus 4.6 (Anthropic, 2026).

6. Subtask 1: Question Interpretation

6.1. Prompting and Constraint Enforcement

All Subtask 1 approaches share a common instruction that explicitly encodes the constraints. We apply deterministic post-processing: (i) strip artifacts (e.g., bullets, quotation marks, prefixes such as "Answer:"), (ii) enforce a trailing "?" if missing, and (iii) enforce the 15-word limit by whitespace tokenization and truncation (consistent with the official evaluation policy).

Decoding uses temperature 0.2, top-p 0.9, max 64 new tokens, and two random seeds.

6.2. Prompting Strategies

Zero-shot prompting. We provide only the constraint instruction and the patient question.

Few-shot prompting. We prepend four exemplars (patient question $\rightarrow$ clinician question) to induce the desired query style.

RAG few-shot prompting. We implement RAG in-context learning by selecting exemplars dynamically for each test input. We embed candidate exemplars with SentenceTransformers (Reimers and Gurevych, 2019) (all-mpnet-base-v2; MPNet backbone (Song et al., 2020)) and build a FAISS inner-product index (Johnson et al., 2017). At inference time, the patient question retrieves top-k similar training inputs; we inject the top exemplars as demonstrations and then generate the clinician question. This approach is closely related to nearest-neighbor prompting (Shi et al., 2022), but adapted to a clinical rewriting format with explicit length constraints.

Clinical retrieval variant. We also test a clinical embedding variant using BioBERT (Lee et al., 2020) to compute dense sentence representations and (optionally) a cross-encoder reranker (details in the appendix).

Closed-LLM baseline. We run Claude Opus 4.5 (Anthropic, 2025) with the same zero-shot and few-shot prompts for a closed vs. open comparison on Subtask 1.

6.3. Supervised Fine-Tuning and Preference Optimization

SFT. We fine-tune Qwen3-30B-A3B-Instruct-2507 and gpt-oss-120b with LoRA adapters (Hu et al., 2022) using chat-formatted inputs.

DPO. We apply DPO (Rafailov et al., 2023) on top of the SFT LLM by sampling candidates per prompt and constructing preference pairs using ROUGE-L as a heuristic selector. (We include this as an exploratory alignment method; see Discussion for limitations.)

7. Subtask 2: Evidence Identification

7.1. Hybrid Similarity-Based Few-Shot Prompting

Example selection. We select k=5 few-shot exemplars per target case using a hybrid similarity measure that combines semantic and lexical signals:

$$\mathrm{sim}(q, q_i) = \alpha \cdot \cos(\mathbf{e}_q, \mathbf{e}_{q_i}) + (1 - \alpha) \cdot \mathrm{RL}(q, q_i) \quad (1)$$

where $\mathbf{e}$ denotes sentence embeddings from all-MiniLM-L6-v2 (Reimers and Gurevych, 2019), RL is the ROUGE-L F-measure (Lin, 2004), and α=0.7. The combination favors semantically similar cases while also rewarding surface-level overlap in clinical terminology, which is important for matching cases within the same clinical specialty. Examples are drawn from the development set, excluding the current target case.

Prompt construction. The prompt consists of: (i) a system instruction defining the evidence selection role, (ii) a task description with five criteria for essential evidence (sentences that directly address the question, contain critical clinical findings, provide necessary diagnostic or treatment context, are indispensable for a complete answer, or include key lab values/medications/procedures), (iii) the k most similar few-shot examples, each presenting the clinician question, numbered note sentences, and gold essential sentence IDs, and (iv) the target case without the answer. The LLM is instructed to respond in a structured format: `Answer: The relevant sentence IDs are: <comma-separated IDs>`.

Decoding and output parsing. All LLMs use greedy decoding (temperature 0.0) for deterministic, reproducible output. A regex-based parser extracts sentence IDs from the structured output, handling variations such as JSON lists, free-text ID mentions, and whitespace differences.

Union ensemble. We combine predictions from GPT-OSS-120B and Qwen3-30B-A3B using a union (logical OR) strategy: a sentence ID is included in the final prediction if either LLM selects it. This design exploits the complementary precision–recall profiles observed during development: GPT-OSS-120B tends toward conservative, high-precision predictions (strict micro-P 71.2% on dev), while Qwen3-30B-A3B favors broader selection with higher recall (strict micro-R 63.6% on dev). Both LLMs are served simultaneously and their prediction sets are merged and sorted by sentence ID.

Closed-LLM baseline. To contextualize the open-LLM ensemble, we run the same hybrid similarity-based 5-shot prompting pipeline (Section 7.1) with Claude Opus 4.6 (Anthropic, 2026) via the Anthropic API. This provides a single-LLM proprietary baseline that uses the identical prompt template, example selection strategy, and output parsing, isolating the effect of LLM choice from all other pipeline components. Comparing the ensemble against this baseline quantifies whether combining two smaller open-weight LLMs can match or exceed a state-of-the-art closed LLM under controlled conditions.

8. Subtask 3: Answer Generation

8.1. Shared Prompt and Post-Processing

All Subtask 3 approaches share a system prompt instructing the LLM to act as a clinical QA assistant with explicit constraints: ≤ 75 words, use only facts from the clinical note excerpt, maintain professional register, and avoid speculation. The user prompt structures the input as three labeled sections (patient question, clinician question, clinical note excerpt).

All outputs undergo deterministic post-processing: (i) removal of `<think>` blocks (common in reasoning-enabled LLMs), (ii) stripping of quotation marks, bullets, and LLM-specific prefixes, (iii) medical term normalization, and (iv) truncation to 75 words with proper sentence-ending punctuation.

8.2. Prompting Strategies

Zero-shot prompting. The LLM receives only the system instruction and the case (no exemplars). We evaluate three backbones: Qwen3-30B-A3B, GPT-OSS-120B, and Claude Opus 4.5. Decoding uses temperature 0.3, top-p 0.9. For local LLMs, two candidates are generated with different seeds and reranked; Claude uses a single seed.

Few-shot prompting. Four fixed exemplars (patient question, clinician question, note excerpt, gold answer) are prepended as user–assistant turn pairs. Exemplars are randomly sampled from the training split (seed 13).

RAG few-shot. Instead of fixed exemplars, we dynamically retrieve the four most relevant training cases per test input from a FAISS inner-product index built with `all-mpnet-base-v2` embeddings (§6.2). The query concatenates the patient and clinician questions; top-4 neighbors are retrieved and used as demonstrations.

Closed-LLM baseline. We run Claude Opus 4.5 (Anthropic, 2025) with the same zero-shot and few-shot prompts for a closed vs. open comparison.

8.3. Supervised Fine-Tuning and Preference Optimization

SFT. We fine-tune Qwen3-30B-A3B-Instruct-2507 with QLoRA (4-bit NF4) on training cases formatted as chat interactions. Two configurations are explored: (i) standard (LoRA r=8, α=16, 5 epochs, max length 1024; best checkpoint at step 35) and (ii) extended (LoRA r=16, α=32, 10 epochs, max length 2048; best checkpoint at step 70). Both use learning rate 2×10^{-5}, effective batch size 16, early stopping on validation loss, and AdamW.

DPO. Starting from the base LLM, we generate multiple candidates per training case at varied temperatures (0.2–0.4), score them against reference answers with ROUGE-L, and select the best/worst as chosen/rejected pairs. DPO training uses β=0.1, LoRA r=8 (standard) or r=16 (extended), and learning rates of 5×10^{-6} and 1×10^{-6} respectively.

9. Subtask 4: Evidence Alignment

9.1. Prompting Strategy

LLM selection. We evaluated three state-of-the-art closed-source LLMs via API on the development set: GPT-5.1 (OpenAI, 2025), GPT-5.2 (OpenAI, 2025), and Claude Opus 4.6 (Anthropic, 2026). Based on development set performance, Claude Opus 4.6 achieved the best results and was selected for final evaluation on the test set.

Zero-shot prompting. We adopt a zero-shot prompting strategy without any task-specific demonstrations. For each answer sentence $a_i \in \mathcal{A}$, we construct a prompt that includes the patient question q, the full clinical note excerpt $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$, and the complete answer $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ as context, and instruct the LLM to identify the supporting evidence sentences from $\mathcal{S}$ for the given a_i. Formally, the LLM is queried as $f(a_i \mid q, \mathcal{S}, \mathcal{A}) \to \mathcal{S}' \subseteq \mathcal{S}$, where $\mathcal{S}'$ is the predicted set of supporting sentences to create a mapping $f : \mathcal{A} \times \mathcal{S} \to \{0, 1\}$ that aligns each answer sentence to its supporting sentence(s) in the clinical note excerpt.

10. Experimental Setup

Metrics. We follow the official evaluation scripts (Organizers, 2026). For Subtask 1, we report ROUGE-L, BERTScore, AlignScore, and MEDCON; the primary score is $(\text{ROUGE-L} + \text{BERTScore})/2$ (Lin, 2004; Zhang et al., 2020; Laban et al., 2022). MEDCON computes clinical concept overlap using UMLS concept extraction (QuickUMLS) (Bodenreider, 2004; Soldaini and Goharian, 2016). For Subtask 2, we report strict and lenient micro/macro precision, recall, and F1. Under strict evaluation, only essential sentences count as gold positives; under lenient evaluation, supplementary sentences additionally do not incur false-positive penalties. For Subtask 3, the official overall score aggregates BLEU, ROUGE-1/2/L/Lsum, SARI, BERTScore, AlignScore, and MEDCON. During development we used a primary score of $(\text{ROUGE-L} + \text{BERTScore})/2$ and BLEU as a secondary metric. For Subtask 4, we report micro-averaged Precision, Recall, and F1 over all predicted alignment links as the primary evaluation metrics. In addition, we report macro-averaged Precision, Recall, and F1 as supplementary analysis.

Implementation notes. We use two random seeds for generation-based runs (Subtask 1) and perform deterministic post-processing to meet format constraints. Subtask 2 uses greedy decoding (temperature 0.0) for deterministic output. Open-weight LLMs are served via vLLM on NVIDIA H100 GPUs; Claude models are accessed via the Anthropic API. Training and adapter configurations are documented in Appendix A.

11. Results

11.1. Subtask 1 Results

We report validation results on 100 released cases (IDs 21–120) using the official scorer. Table 1 summarizes our main Subtask 1 approaches. On the development set, RAG few-shot prompting achieves the highest primary score (31.05) and overall score (25.61), followed by the clinical retrieval variant (25.41) and Claude few-shot (25.27). DPO achieves the highest AlignScore (18.37) and

Approach	Primary	R-L	BERT	Align	MEDCON	Overall
RAG few-shot	**31.05**	**23.63**	38.46	14.91	25.45	**25.61**
RAG few-shot (clinical)	30.78	23.34	38.22	15.18	24.92	25.41
Claude few-shot	30.84	22.84	**38.84**	14.90	24.51	25.27
Qwen SFT	27.61	18.79	36.44	17.76	24.11	24.27
Qwen DPO (full)	24.53	16.44	32.61	**18.37**	**26.58**	23.50

Table 1: Validation scores for Subtask 1 on cases 21–120 (100 cases). Primary = (ROUGE-L + BERTScore)/2 on 0–100 scale. Overall = official mean of AlignScore, ROUGE-L, MEDCON, and BERTScore. MEDCON uses UMLS concept extraction (QuickUMLS) with UMLS knowledge sources.

Approach	R-L	BERT	Align	MC	OA
Claude fs	**21.76**	**38.01**	**21.71**	**26.28**	**26.94**
RAG fs	18.46	33.49	14.39	18.99	21.33
Qwen DPO	10.68	20.55	16.77	20.83	17.21

Table 2: Codabench Subtask 1 test scores (cases 121–167). MC = MEDCON, OA = overall score.

#	Team	OA
1	HealthNLP_Retrievers	31.2
2	KPSCMI	30.8
3	OptiMed	29.9
4	Neural	28.9
5	Yale-DM-Lab	27.1
6	**Ours**	**26.9**

Table 3: Subtask 1 Codabench leaderboard (6/13 teams). OA = official overall score.

MEDCON (26.58) but the lowest primary and overall scores.

Test set (cases 121–167). The official test labels are withheld; scores are returned via Codabench. Our three Subtask 1 submissions are Claude few-shot, RAG few-shot, and Qwen DPO (full-train). Table 2 lists the official test scores; Table 3 provides leaderboard context.

On the test set, the ranking reverses substantially: Claude few-shot achieves the best overall score among our submissions (26.94), outperforming RAG few-shot (21.33) by 5.6 points. Claude few-shot leads on all individual metrics, with particularly large margins on BERTScore (38.01 vs. 33.49) and MEDCON (26.28 vs. 18.99). Qwen DPO trails further (17.21). Our best submission (Claude few-shot) places 6th on the official leaderboard, 4.3 points behind the top-ranked system (HealthNLP_Retrievers, 31.2).

11.2. Subtask 2 Results

We submitted three configurations to the official Codabench leaderboard: (1) Claude Opus 4.6 with 5-shot prompting, (2) the union ensemble of GPT-OSS-120B and Qwen3-30B-A3B with 5-shot prompting, and (3) the same ensemble in a zero-shot setting (no in-context examples). Table 4 reports the official test scores.

Claude Opus 4.6 achieves the highest strict micro-F1 among our submissions (58.8), placing 10th on the leaderboard out of 14 teams. The open-LLM ensemble with few-shot prompting reaches 56.7 (12th), trailing Opus by only 2.1 F1 points while achieving higher recall (53.6 vs. 52.3). Removing few-shot examples (zero-shot ensemble) drops performance substantially to 51.4, confirming the value of in-context demonstrations for calibrating evidence selection.

All three runs exhibit notably high lenient precision (up to 84.2% for Opus), indicating that sentences classified as false positives under strict evaluation are predominantly supplementary rather than not-relevant—the LLMs select clinically pertinent sentences even when they do not meet the strict essential threshold.

11.3. Subtask 3 Results

We evaluated 10 configurations on the 20-case development set, covering four paradigms (zero-shot, few-shot, RAG few-shot, SFT/DPO), and submitted three systems to the official test phase. Table 5 summarizes development results for representative approaches; Table 6 reports official test scores.

Development set results. RAG few-shot with Claude Opus 4.5 achieves the highest primary score (60.5), BLEU (12.7), and MEDCON (54.6) on the 20-case development set (Table 5) and outperforms the next-best configuration (RAG few-shot Qwen) by 2.4 primary points. Dynamically retrieved exemplars consistently outperform fixed few-shot demonstrations across both LLM backbones (Claude: 60.5 vs. 57.8; Qwen: 58.1 vs. 55.1).

Effect of fine-tuning. Extended SFT training improves over the standard configuration by 2.2 primary points (57.2 vs. 55.0). Both SFT variants trail RAG few-shot Qwen (58.1). DPO does not improve over SFT (53.3 vs. 55.0).

		Strict Micro				Lenient Micro			
Run	OA	P	R	F1	Macro F1	P	R	F1	Macro F1
Opus 4.6 (5-shot)	**58.8**	**67.1**	52.3	**58.8**	**58.1**	**84.2**	52.3	**64.5**	**63.6**
Ensemble (5-shot)	56.7	60.2	**53.6**	56.7	55.1	74.7	**53.6**	62.4	60.4
Ensemble (0-shot)	51.4	61.8	44.0	51.4	50.3	76.7	44.0	55.9	54.6
Leaderboard context (top-3 teams):									
1st (Neural)	63.7	60.2	67.6	63.7	64.8	77.8	67.6	72.4	73.1
2nd (OptiMed)	63.2	56.7	71.3	63.2	64.1	73.1	71.3	72.2	72.9
3rd (UIC-AIHealth4All)	62.9	59.3	67.0	62.9	63.0	74.3	67.0	70.4	70.5

Table 4: Subtask 2 evidence identification results on the official Codabench test set (14 participating teams). OA = Overall (Strict Micro F1), the official ranking metric. Ensemble = union of GPT-OSS-120B and Qwen3-30B-A3B. Macro F1 = strict macro F1 for our runs; shown for comparison. Best values among our submissions in bold.

Approach	Primary	BLEU	MC
RAG fs Claude	**60.5**	**12.7**	**54.6**
RAG fs Qwen	58.1	9.1	46.7
Few-shot Claude	57.8	9.6	46.9
SFT-long Qwen	57.2	8.3	39.0
Few-shot Qwen	55.1	6.3	42.4
SFT Qwen	55.0	4.5	34.5
Zero-shot Claude	54.1	6.3	43.1
DPO Qwen	53.3	4.2	35.7
Zero-shot Qwen	53.0	4.0	35.6
RAG fs OSS	52.3	5.8	46.0
Few-shot OSS	52.3	5.7	43.2
Zero-shot OSS	51.4	4.9	41.3

Table 5: Subtask 3 development results (20 cases; selected configurations). Primary = (ROUGE-L + BERTScore)/2 $\times$ 100. MC = MEDCON.

Backbone comparison. Under identical prompting, a consistent ranking emerges across backbones. With RAG few-shot, Claude Opus 4.5 outperforms Qwen3-30B-A3B by 2.4 primary points and 7.9 MEDCON points; under zero-shot, the MEDCON gap widens further (43.1 vs. 35.6). GPT-OSS-120B consistently trails both Claude and Qwen and records systematically lower primary scores.

Test set results and leaderboard placement. RAG few-shot Claude achieves an overall score of 36.33 on the official test set (Table 6) and ranks **first** out of 13 participating teams (Table 7). The margin over the second-ranked team (TAMU-NLP-Lab, 36.2) is narrow (Δ=0.1). RAG few-shot Claude leads on all individual metrics except SARI, where SFT-long scores highest (59.31 vs. 58.60).

RAG gain over zero-shot. On the test set, RAG few-shot Claude outperforms zero-shot Claude by 3.5 overall points (36.33 vs. 32.83). The largest margins appear on BLEU (9.94 vs. 5.28) and BERTScore (46.79 vs. 41.76). The MEDCON gap is smaller (43.14 vs. 41.47).

11.4. Subtask 4 Results

We first present a comparative analysis of the LLMs described in Section 9.1 on the development set under both the w/ and w/o patient question settings, followed by the evaluation of the best-performing LLM on the test set.

Comparison across LLMs. As shown in the lower half of Table 8, Claude Opus 4.6 achieves the best overall performance across all metrics, attaining a micro F1 of 0.975 and a macro F1 of 0.973 under the w/ patient question setting, substantially outperforming both GPT variants. GPT-5.2 consistently ranks second, with micro F1 scores of 0.868 and 0.878 in the two settings, while GPT-5.1 records the lowest scores across all metrics. Notably, all three LLMs achieve consistently high recall, suggesting that LLMs are generally capable of retrieving relevant evidence sentences; however, precision varies more markedly across LLMs, with Opus 4.6 demonstrating a considerably stronger ability to suppress spurious alignments.

Effect of patient question. As shown in the upper half of Table 8, incorporating the patient question q into the input context consistently improves performance across all LLMs and metrics. The most substantial gains are observed for Claude Opus 4.6, where micro Precision increases from 0.783 to 0.958 and micro F1 improves from 0.875 to 0.975, indicating that the patient question provides critical contextual signal for disambiguating relevant evidence sentences. GPT-5.1 and GPT-5.2 also benefit from the inclusion of q, with moderate gains in precision and F1 while recall remains relatively stable. These results highlight the importance of the patient question as contextual grounding, particularly for more capable LLMs that can better exploit

Run	OA	Align	BERT	BLEU	MC	R-L	SARI
RAG fs Claude	**36.33**	**31.71**	**46.79**	**9.94**	**43.14**	**27.81**	58.60
SFT-long (Qwen, local)	34.16	29.86	43.56	8.36	37.53	26.30	**59.31**
Zero-shot Claude	32.83	28.08	41.76	5.28	41.47	23.12	57.28

Table 6: Subtask 3 official test results. OA = overall score (composite of all metrics). MC = MEDCON. Best values in bold.

#	Team	OA
1	**Ours**	**36.3**
2	TAMU-NLP-Lab	36.2
3	BIT.UA-AAUBS	35.6
4	Neural	35.2
5	HealthNLP_Retrievers	34.6
Median (13 teams)		32.3

Table 7: Subtask 3 Codabench leaderboard (top-5 of 13 teams). OA = official overall score.

Run	P_{mi}	R_{mi}	$F1_{mi}$	P_{ma}	R_{ma}	$F1_{ma}$
W/O Patient Question						
GPT-5.1	0.735	0.884	0.803	0.752	0.887	0.797
GPT-5.2	0.799	<u>0.949</u>	0.868	0.847	0.947	0.882
Opus-4.6	0.783	**0.993**	0.875	0.812	**0.993**	0.876
W/ Patient Question						
GPT-5.1	0.774	0.891	0.835	0.811	0.894	0.835
GPT-5.2	<u>0.827</u>	0.935	<u>0.878</u>	<u>0.878</u>	0.937	<u>0.896</u>
Opus-4.6	**0.958**	**0.993**	**0.975**	**0.962**	<u>0.990</u>	**0.973**

Table 8: Subtask 4 answer–evidence alignment results with and without the patient question q as input context. Subscripts $_{mi}$ and $_{ma}$ denote micro- and macro-averaged metrics, respectively. **Bold** indicates the best score and <u>underline</u> indicates the second-best score in each column.

the additional information.

Test Set Results. We submitted the predictions of Claude Opus 4.6 with patient question included for the final evaluation on the test set. The system achieves a micro F1 of 81.27 and a macro F1 of 82.09, with a micro Precision of 86.93, micro Recall of 76.31, macro Precision of 87.54, and macro Recall of 79.41. Compared to the development set results, there is a noticeable drop in both precision and recall, which is expected due to the distribution shift between the development and test sets. Nevertheless, the system maintains relatively strong precision, suggesting that the predicted alignment links are generally reliable.

12. Discussion

Subtask 1 performance trends and test-time shift. On cases 21–120, RAG few-shot prompting is consistently strongest, with the clinical retrieval variant close behind. The Claude few-shot baseline is competitive and outperforms open baselines on several metrics. SFT improves over prompt-only baselines but trails RAG on the official overall score. DPO increases AlignScore and MEDCON but does not translate into higher overall performance. On the test set (cases 121–167), the ranking reverses substantially: Claude few-shot leads by 5.6 overall points (26.94 vs. 21.33), whereas RAG few-shot led on development by only 0.34 points. This highlights a significant distribution shift between the released labeled cases and the hidden test set. RAG few-shot retrieves exemplars from the training pool using semantic similarity; when test inputs diverge from this pool, the retrieved demonstrations may be poorly matched, amplifying template-like phrasing and reducing lexical as well as semantic overlap with the references. In contrast, the Claude backbone's stronger language modeling generalizes more robustly without reliance on exemplar quality. The large drop in RAG few-shot MEDCON (25.45 on dev vs. 18.99 on test) suggests that retrieved exemplars may also bias clinical terminology toward training-set conventions that do not transfer. The DPO configuration degrades further on the test set (17.21), confirming that ROUGE-L-based preference optimization does not align with the official multi-metric objective, particularly under distribution shift.

Subtask 2: Open-LLM ensemble vs. proprietary baseline. Claude Opus 4.6 achieves the highest strict micro-F1 among our submissions (58.8), driven by substantially higher precision (67.1 vs. 60.2). However, the union ensemble of two open-weight LLMs achieves higher recall (53.6 vs. 52.3) and remains competitive on F1 (56.7 vs. 58.8). This demonstrates that complementary open LLMs—one exhibiting higher precision (GPT-OSS-120B) and one exhibiting higher recall (Qwen3-30B-A3B) in our experiments—can approach proprietary-LLM performance while preserving full data sovereignty, which is particularly relevant in clinical settings where note text may be subject to data use agreements. A gap of approximately 5 F1 points to the

top-ranked system (Neural, 63.7) suggests room for improvement, potentially through threshold tuning, additional ensemble members, or fine-tuning on evidence labels.

Impact of few-shot exemplars on evidence selection. Comparing the ensemble in few-shot and zero-shot settings reveals a consistent benefit of in-context demonstrations: 5-shot prompting improves strict micro-F1 by 5.3 points (56.7 vs. 51.4) and recall by 9.6 points (53.6 vs. 44.0). The hybrid similarity-based selection (Eq. 1) retrieves cases from similar clinical specialties, providing the LLM with domain-appropriate evidence patterns. This confirms that example selection substantially calibrates the LLM's evidence threshold, reducing both under-selection (zero-shot conservatism) and over-selection.

Strict vs. lenient evaluation gap. All Subtask 2 runs exhibit a large gap between strict and lenient precision (e.g., 67.1 vs. 84.2 for Opus), indicating that the majority of strict false positives are supplementary sentences rather than not-relevant ones. From a clinical utility perspective, this is encouraging: the LLMs over-predict somewhat but predominantly select sentences that are still clinically pertinent.

Subtask 3: RAG gain over instruction-only prompting. Our RAG few-shot Claude submission ranks first out of 13 teams on the official leaderboard (36.3), narrowly ahead of TAMU-NLP-Lab (36.2) and BIT.UA-AAUBS (35.6). On the test set, RAG few-shot Claude outperforms zero-shot Claude by 3.5 overall points, with the largest gaps on BLEU and BERTScore. The MEDCON gap is notably smaller, which indicates that zero-shot Claude already captures relevant clinical concepts but produces surface-level phrasing that deviates from reference conventions. Dynamically retrieved exemplars therefore primarily guide answer style and structure rather than clinical content.

Local fine-tuning as a privacy-preserving alternative. The locally fine-tuned SFT-long submission achieves 94% of the top score without API access and thus represents a viable option for locally hosted clinical QA under data-privacy constraints. DPO did not improve over SFT on the development set, likely because ROUGE-L provides too coarse a preference signal to capture the nuanced quality that the multi-metric evaluation rewards. This finding aligns with the Subtask 1 observation that ROUGE-L-based preference optimization does not transfer reliably to composite evaluation objectives.

LLM quality vs. LLM size. Under identical prompting, Claude Opus 4.5 consistently outperforms both open-weight LLMs. More notably, the Qwen3-30B-A3B MoE architecture outperforms GPT-OSS-120B despite fewer active parameters. GPT-OSS records systematically lower primary scores, which points to weaker semantic alignment with clinical reference answers. Raw parameter count alone is therefore a poor predictor of clinical text generation quality, and architecture choice as well as training data composition likely play a more decisive role.

Implications for downstream grounding. Subtask 1 errors propagate: an overly generic clinician query may reduce evidence recall (Subtask 2), while an overly specific query may bias selection toward irrelevant details. We therefore recommend that Subtask 2 systems (i) optionally incorporate both patient and clinician questions as queries and (ii) include conservative fallbacks for ambiguous interpretation.

13. Conclusion

We described a modular system for four ArchEHR-QA subtasks, centered on RAG prompting and constraint-aware generation. For Subtask 1, RAG few-shot prompting yields the strongest development performance, but Claude few-shot proves substantially more robust on the hidden test set (26.94 vs. 21.33 overall), indicating that retrieval-based exemplar selection is sensitive to distribution shift while a stronger LLM backbone generalizes more reliably. For Subtask 2, Claude Opus 4.6 achieves the best overall F1 (58.8), while a union ensemble of two open-weight LLMs attains competitive performance (56.7) with higher recall, offering a viable open-source alternative under data-privacy constraints. For Subtask 3, we systematically compared 13 configurations across prompting, RAG, SFT, and DPO paradigms. RAG few-shot with Claude Opus 4.5 achieves the highest official test score (36.33), ranking first out of 13 teams on the Codabench leaderboard. RAG provides the largest individual gain; local SFT with QLoRA achieves 94% of the top score while preserving full data sovereignty. For Subtask 4, we employ Claude Opus 4.6 (API) in a prompt-only, zero-shot configuration to identify and extract the EHR sentences that support the given response sentence. This approach achieves an official micro-F1 score of 81.3, ranking 2nd out of 16 participating teams.

14. Limitations

Automatic metrics. Our evaluation relies on automatic metrics (ROUGE, BERTScore, AlignScore, MEDCON for Subtask 1; precision/recall/F1 for Subtask 2) and lacks clinician review; correlations with real clinical utility may be imperfect.

Data scale and shift. The released labeled split used for Subtask 1 development is limited, which can amplify variance. For Subtask 2, the development set of 20 cases also serves as the few-shot example pool, which may limit example diversity. For Subtask 3, the same 20-case dev set yields wide metric confidence intervals. Distribution shift between released data and the hidden test set may substantially affect retrieval-based and preference-optimized systems.

Closed-LLM dependence. While we report a strong closed-LLM baseline for both Subtask 1 (Claude Opus 4.5) and Subtask 2 (Claude Opus 4.6), reproducibility and cost considerations may limit their use in practice. The open-LLM ensemble addresses this partially but still relies on large-scale infrastructure for serving 120B-parameter LLMs.

15. Ethics Statement

This work uses de-identified clinical data distributed via PhysioNet under an approved data use agreement. All experiments were conducted on data that had undergone the MIMIC de-identification pipeline; no attempt was made to re-identify patients, and no individual patient data is reported in this paper.

Our system is designed as a research prototype for a shared task and is not intended for direct clinical deployment. Automatically generated answers to patient questions carry a risk of factual errors, hallucinated clinical content, or misleading omissions that could cause patient harm if acted upon without clinician review. We therefore emphasize that any downstream application of these methods must include clinician-in-the-loop verification before information is communicated to patients.

The use of proprietary LLM APIs (Claude, GPT) for processing clinical text raises data governance concerns, as patient data may traverse third-party infrastructure. We address this partially by evaluating open-weight alternatives that can be hosted locally, preserving data sovereignty. We encourage future work to prioritize locally deployable models for clinical NLP applications where data use agreements restrict external data transfer.

The computational cost of serving large language models (up to 120B parameters on H100 GPUs) has environmental implications. We report all model sizes and hardware configurations to support reproducibility and enable informed comparisons of the accuracy–efficiency trade-off.

16. Bibliographical References

Anthropic. 2025. Claude opus 4.5 system card. https://www.anthropic.com/claude-opus-4-5-system-card. Accessed: 2026-01-23.

Anthropic. 2026. Claude opus 4.6 system card. https://www.anthropic.com/claude-opus-4-6-system-card. Accessed: 2026-03-02.

Anthropic. 2026. Introducing Claude Opus 4.6. Technical report, Anthropic.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hugging Face. 2024. Trl: Transformer reinforcement learning. https://huggingface.co/docs/trl/. Accessed: 2026-01-23.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Philippe Laban, Chien-Sheng Xu, Rahul Aralikatte, Avinash Thawani, Alexander R. Fabbri, Trevor Cohn, and Marti A. Hearst. 2022. AlignScore: Evaluating factual consistency with a unified entailment framework. *arXiv*.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

OpenAI. 2025. GPT-5.1: A smarter, more conversational ChatGPT. Technical report, OpenAI.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv*.

OpenAI. 2025. Introducing GPT-5.2. Technical report, OpenAI.

ArchEHR-QA Organizers. 2026. ArchEHR-QA @ LREC 2026 shared task: Grounded question answering from electronic health records. https://archehr-qa.github.io/. Accessed: 2026-01-23.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv*.

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 and the 9th International Joint Conference on Natural Language Processing*.

Weijia Shi, Ruiqi Gao, Yue Zhang, Yiming Chen, Jie Zhou, Xuanjing Huang, and Furu Wei. 2022. kNN-prompt: Nearest neighbor zero-shot inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Luca Soldaini and Nazli Goharian. 2016. QuickUMLS: a fast, unsupervised approach for medical concept extraction. In *Proceedings of the 2016 Workshop on Biomedical Natural Language Processing*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv*.

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

17. Language Resource References

Asma Ben Abacha and Dina Demner-Fushman. 2019. *MeQSum: A Corpus of Medical Question Summarization*. U.S. National Library of Medicine. PID https://github.com/abachaa/MeQSum.

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 Clinical Database*. PhysioNet, 1.4. PID https://doi.org/10.13026/C2XW26.

Sarvesh Soni and Dina Demner-Fushman. 2026. *ArchEHR-QA: A Dataset for Addressing Patient's Information Needs related to Clinical Course of Hospitalization*. PhysioNet. PID https://doi.org/10.1038/s41597-026-06639-z.

U.S. National Library of Medicine. 2024. *UMLS – Unified Medical Language System*. U.S. National Library of Medicine. PID https://www.nlm.nih.gov/research/umls/.

A. Training Setup

We train LoRA adapters with TRL's SFTTrainer and DPOTrainer (Hugging Face, 2024). Unless otherwise noted, runs use bf16, gradient checkpointing, AdamW, and sequence length 512. Full-train runs (for test submissions) use all available labeled pairs without a dev split and omit evaluation during training. Training is executed sequentially and logged per run under `runs/<approach>/launcher.log`.

- SFT: learning rate 2×10^{-5}, 1 epoch, batch size 1, gradient accumulation 4.
- LoRA (Qwen): $r = 64$, $\alpha = 128$, dropout 0.05; target modules `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`.
- LoRA (GPT-OSS): $r = 16$, $\alpha = 64$, dropout 0.05.
- DPO: $\beta = 0.1$, 2 candidates per prompt; full-train uses all rows.
- Distillation (Subtask 1 exploratory): Claude pseudo labels merged with ArchEHR + MeQSum.
- Retrieval: top-k=6, 4 exemplars; embeddings from SentenceTransformers; clinical variant uses BioBERT.
- Hardware: NVIDIA H100 80GB GPUs.

B. Run Inventory (Subtask 1)

Table 9 lists all Subtask 1 configurations evaluated during development and submitted to the official test phase. Prompting-based runs (zero-shot and few-shot) serve as baselines; retrieval-augmented variants build on these by replacing fixed exemplars with dynamically retrieved nearest neighbors. Supervised fine-tuning and DPO runs use LoRA adapters trained on the released labeled pairs, with the full-train DPO variant submitted to the test set. All runs share the same post-processing pipeline to enforce the 15-word output constraint.

C. Run Inventory (Subtask 2)

Table 10 documents the three configurations submitted for Subtask 2. All runs use the same hybrid similarity-based example selection strategy and identical prompt template, isolating the effect of model choice and the presence or absence of in-context demonstrations. The ensemble runs combine GPT-OSS-120B and Qwen3-30B-A3B via a union merge; the Claude Opus 4.6 run serves as a single-model proprietary reference under otherwise identical conditions.

D. Run Inventory (Subtask 3)

Table 11 provides the full run inventory for Subtask 3, covering all 10 configurations evaluated on the development set and the three submitted to the official test phase (marked in bold). The inventory spans four paradigms: zero-shot prompting, fixed few-shot prompting, retrieval-augmented few-shot prompting, and parameter-efficient fine-tuning via QLoRA, with an additional DPO variant. GPT-OSS-120B variants follow the same prompting templates as the corresponding Qwen runs and are omitted from the table for brevity. All runs apply identical 75-word post-processing and use gold clinician questions as input.

E. Leaderboards

Table 12 places this submission in the context of the official Codabench leaderboard, where our system ranked second out of 16 participating teams with a micro-F1 of 81.3.

Approach	Data	LLM	Training/Inference	Notes
prompt_zeroshot	Released labeled inputs (21–120)	Qwen3-30B-A3B	temp 0.2, top-p 0.9, 2 seeds	base prompt
prompt_fewshot	Released labeled inputs (21–120)	Qwen3-30B-A3B	temp 0.2, top-p 0.9, 2 seeds	4 exemplars from training pairs
prompt_zeroshot_claude	Released labeled inputs (21–120)	Claude Opus 4.5	temp 0.2, 2 seeds	closed LLM baseline
prompt_fewshot_claude	Released labeled inputs (21–120)	Claude Opus 4.5	temp 0.2, 2 seeds	4 exemplars
rag_fewshot	Released labeled inputs (21–120)	Qwen3-30B-A3B	temp 0.2, top-p 0.9	FAISS all-mpnet; top-k 6
rag_fewshot_clinical	Released labeled inputs (21–120)	Qwen3-30B-A3B	temp 0.2, top-p 0.9	BioBERT embed + rerank
sft_qwen_instruct	Train pairs + aug	Qwen3-30B-A3B	LoRA r64, lr 2e-5, ep1	dev split (case-id)
sft_gptoss	Train pairs + aug	gpt-oss-120b	LoRA r16, lr 2e-5, ep1	chat template
sft_qwen_distill	Distill pairs	Qwen3-30B-A3B	LoRA r64, lr 2e-5, ep1	pseudo labels
dpo_qwen_fulltrain	All pairs	Qwen3-30B-A3B	beta 0.1, 2 candidates	full-train for test

Table 9: Run inventory for Subtask 1. Bold approaches were submitted to the official test phase. All runs use max 64 new tokens and post-processing to enforce ≤ 15 words.

Approach	Data	LLM	Inference	Notes
opus_5shot	Dev (20 cases) as few-shot pool; test split for submission	Claude Opus 4.6	temp 0.0, greedy decoding, Anthropic API	5 exemplars via hybrid similarity (α=0.7); gold clinician questions
ensemble_5shot	Dev (20 cases) as few-shot pool; test split for submission	GPT-OSS-120B + Qwen3-30B-A3B (union)	temp 0.0, greedy decoding, vLLM	5 exemplars via hybrid similarity; None-elimination for GPT-OSS; union merge
ensemble_0shot	Test split for submission	GPT-OSS-120B + Qwen3-30B-A3B (union)	temp 0.0, greedy decoding, vLLM	No in-context examples; None-elimination for GPT-OSS; union merge

Table 10: Run inventory for Subtask 2. Hybrid similarity combines all-MiniLM-L6-v2 embeddings (weight 0.7) and ROUGE-L (weight 0.3). All runs use gold clinician questions and regex-based output parsing.

Approach	Data	LLM	Inference	Notes
prompt_zeroshot	Train (gold answers)	Qwen3-30B-A3B	temp 0.3, top-p 0.9, 2 seeds, rerank	Zero-shot; 75-word post-processing
prompt_fewshot	Train (gold answers)	Qwen3-30B-A3B	temp 0.3, top-p 0.9, 2 seeds, rerank	4 fixed exemplars (seed 13)
rag_fewshot	Train (gold answers)	Qwen3-30B-A3B	temp 0.3, top-p 0.9, 2 seeds, rerank	FAISS mpnet top-6→4
prompt_zeroshot_claude	Train (gold answers)	Claude Opus 4.5	temp 0.3, 1 seed, API	**Submitted to test**
prompt_fewshot_claude	Train (gold answers)	Claude Opus 4.5	temp 0.3, 1 seed, API	4 fixed exemplars
rag_fewshot_claude	Train (gold answers)	Claude Opus 4.5	temp 0.3, 2 seeds, API	**Submitted to test**; FAISS mpnet
sft_qwen_instruct	Train (gold answers)	Qwen3-30B-A3B	QLoRA 4-bit; LoRA r=8, 5 ep	Best ckpt step 35
sft_qwen_instruct_long	Train (gold answers)	Qwen3-30B-A3B	QLoRA 4-bit; LoRA r=16, 10 ep	**Submitted to test**; best ckpt step 70
dpo_qwen	Train (gold answers)	Qwen3-30B-A3B	DPO β=0.1; LoRA r=8, 3 ep	ROUGE-L preference pairs
dpo_qwen_full	Train (gold answers)	Qwen3-30B-A3B	DPO β=0.1; LoRA r=16, 5 ep	Merged adapter at inference

Table 11: Run inventory for Subtask 3. GPT-OSS-120B variants (zero-shot, few-shot, RAG) follow the same template as the corresponding Qwen runs. All runs use gold clinician questions and 75-word post-processing.

#	Team	P_{mi}	R_{mi}	$F1_{mi}$
1	BIT.UA-AAUBS	**88.0**	75.9	**81.5**
2	Ours	86.9	76.3	81.3
3	Yale-DM-Lab	83.3	77.7	80.4
4	OptiMed	80.7	**79.8**	80.3
5	UIC-AIHealth4All	83.6	76.3	79.8

Table 12: Subtask 4 Codabench leaderboard (top-5 of 16 teams). Subscripts $_{mi}$ denote micro-averaged metrics.