From Writing to Asking: Dual-Stage Clinical Note Generation with Active Questioning and Preference Alignment

Anonymous ACL submission

Abstract

Large language models are increasingly used to support clinical documentation generation, but real clinical situations involve heavy documentation workloads and records of uneven quality, where missing key evidence may lead to inaccurate diagnoses. We introduce **Ask–then–Regenerate**[1], a two-stage framework for clinical documentation generation that covers four clinical tasks: admission diagnosis generation, differential diagnosis discussion, treatment planning, and active clinical questioning. In Stage I, a unified Clinical Documentation Generator is trained for the three documentation tasks, with an interaction-guided sampling rule derived from task interaction analysis. In Stage II, the model is further trained to ask targeted questions from preliminary diagnosis and reasoning drafts, and its questioning behavior is aligned with physician preferences using Direct Preference Optimization. The recovered information is then used to regenerate the documentation. Experiments show that Stage I achieves 78.61% average Entity Recall, outperforming zero-shot Qwen3-4B by 9.50 points. Under a Masked-Field Protocol that removes clinically important fragments to simulate incomplete records, Ask–then–Regenerate recovers 52.71% of the average Entity Recall loss, while physician evaluation confirms that DPO improves the clinical value of generated questions.

1 Introduction

Large language models are increasingly being used to support clinical work, including medical question answering, decision support, and clinical note generation, as shown by systems such as HuatuoGPT-II (Chen et al., 2024), Med-PaLM 2 (Singhal et al., 2023), and MedCare (Liao et al., 2024).During patient admission, physicians must convert long and fragmented admission records into diagnoses, reasoning discussions, and treatment plans. This workload is heavier in specialties such as respiratory medicine, where records often contain complex symptoms, imaging findings, laboratory tests, treatment histories, and comorbidities. Although recent medical LLMs can generate fluent clinical text, most existing systems still treat documentation as a passive, single-pass task: given a record, generate the note.

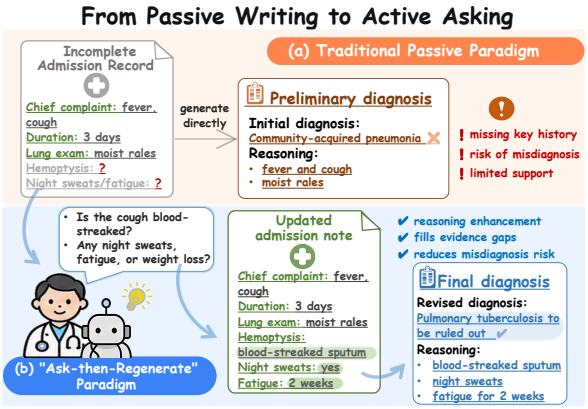

Figure 1: Motivation of Ask–then–Regenerate: asking targeted questions before generating clinical notes.

This setting overlooks a common problem in real admission records. Records are often written under time pressure and may omit key evidence such as symptom duration, exposure history, medication dosage, prior treatment response, risk factors, or pertinent negatives. When the input is incomplete, a model may still produce a fluent note, but the diagnosis, reasoning, or treatment plan can be based on insufficient evidence. Experienced physicians usually do not write final documentation directly from such low-quality records; they first confirm key missing information and then complete the downstream clinical writing. This motivates us to view clinical note generation not only as text generation, but also as active information seeking.

[1]Code is available at https://anonymous.4open.science/r/Ask-then-Regenerate.

We instantiate this view as Ask-then-Regenerate, a framework for active information seeking in clinical documentation. It contains two main components. The first is a Clinical Documentation Generator, a unified model trained on three respiratory medicine documentation tasks: admission diagnosis generation, differential diagnosis discussion, and treatment planning. Since these tasks are related but stylistically distinct, we analyze their pairwise interactions and derive a lightweight interaction-guided sampling prior for stable multi-task fine-tuning. The second stage trains the model to ask targeted questions based on the admission record and preliminary drafts. Given an admission record and the generator's preliminary drafts, this module asks targeted questions to recover missing or weakly supported clinical evidence. The obtained answers are then inserted back into the record, and the generator regenerates the final documentation from the augmented input. To make the questioning behavior clinically useful, we further align this module with physician preferences using Direct Preference Optimization, favoring questions that are informative and likely to recover documentation-relevant evidence.

To quantify how much active questioning improves documentation under incomplete inputs, we evaluate Ask-then-Regenerate in a closed-loop setting. We propose a Masked-Field Protocol that removes clinically important fragments from complete held-out records and returns them only when the model asks for the corresponding missing information. This tests whether the system can recover evidence that matters for downstream documentation, rather than merely generate plausible questions. We compare direct generation from incomplete records, Ask-then-Regenerate with question-conditioned recovery, and generation from complete records, and report Gap Closure as the fraction of missing-information loss recovered by the closed-loop workflow. Our contributions are:

- We propose an Ask–then–Regenerate workflow for clinical documentation under incomplete inputs, with quantitative validation;
- We train draft-guided active questioning with physician preference alignment via DPO;
- We present an empirical analysis of task interactions in clinical multi-task fine-tuning and derive a simple transfer-aware sampling rule;
- We build a deployable respiratory documentation system validated for clinical feasibility.

2 Related Work

Medical LLMs. Recent medical LLMs such as Med-PaLM 2 (Singhal et al., 2023), GPT-4 (Nori et al., 2023), HuatuoGPT-II (Chen et al., 2024), and MedCare (Liao et al., 2024) have shown strong performance in medical question answering and clinical text generation. However, these systems usually operate as passive respondents: given an input, they directly produce an answer or note. In contrast, our work focuses on incomplete admission records and introduces active questioning before final documentation generation.

Medical dialogue systems. Medical dialogue systems such as BianQue (Chen et al., 2023), MedDG (Liu et al., 2022), and DISC-MedLLM (Bao et al., 2023) study multi-turn patient-facing interaction. Earlier work also framed diagnosis as task-oriented dialogue (Wei et al., 2018), and a related line generates clinical notes from doctor-patient conversations (Krishna et al., 2021; Zhang et al., 2021).Our setting differs in both objective and evaluation. We do not aim to optimize dialogue quality itself; instead, questions are generated to recover missing clinical evidence and improve downstream admission diagnoses, reasoning discussions, and treatment plans.

Multi-task learning. Multi-task learning is a long-studied paradigm (Caruana, 1997) and underlies unified text-to-text models (Raffel et al., 2023); it has been studied through data-centric methods such as DoReMi (Xie et al., 2023) and RegMix (Liu et al., 2025), as well as gradient-based methods such as PCGrad (Yu et al., 2020). In clinical NLP, prior work has mainly focused on extraction, classification, or benchmark-style tasks. We instead analyze task interactions among clinical generation tasks and use the observed transfer pattern to guide multi-task fine-tuning.

Preference optimization in healthcare. Preference optimization methods, including RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023), have been used to improve the safety and quality of medical LLM responses (Chen et al., 2024; Liao et al., 2024). Most existing work aligns final answers. Our work applies DPO to active clinical questioning, encouraging questions that physi-

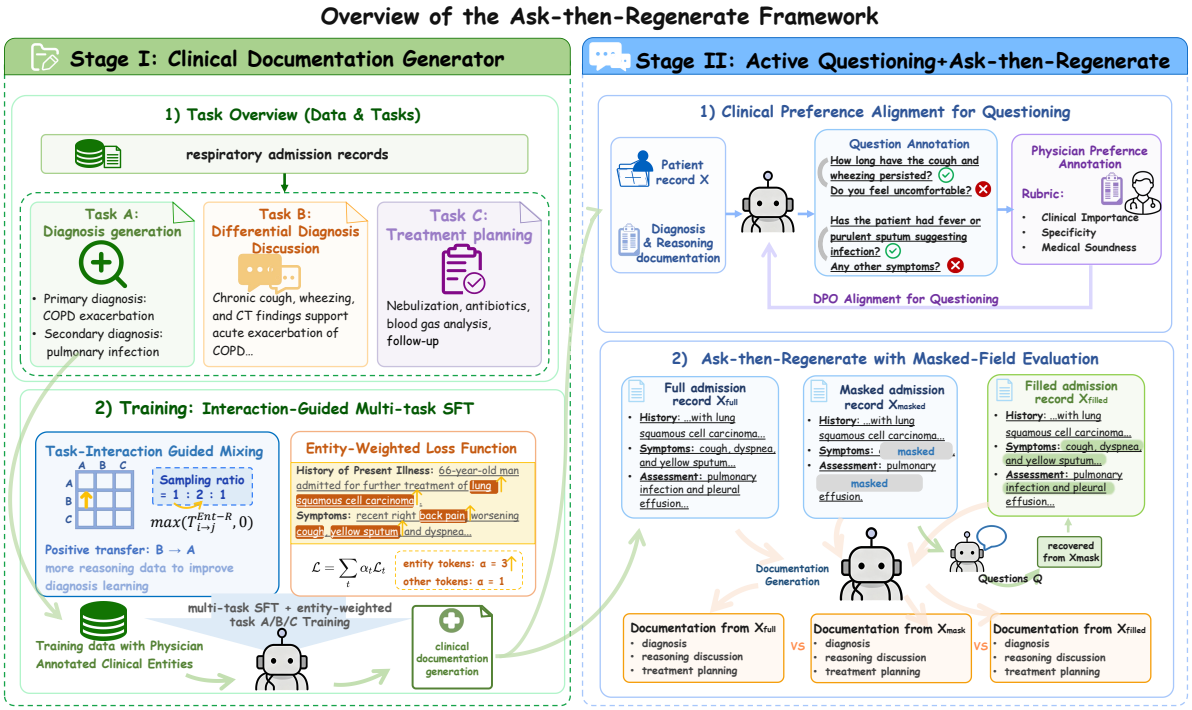

Figure 2: System overview of the proposed Ask-then-Regenerate framework.

cians judge to be important, specific, and medically sound.

3 Method

3.1 System Overview: The Ask–then–Regenerate Workflow

Figure 2 gives the overall architecture of our clinical decision-support workflow. The system is designed to assist diagnosis and treatment planning from respiratory admission records, especially when the available evidence is incomplete or under-documented. Rather than treating clinical note generation as passive one-shot writing, we model it as active information seeking: the model first forms a provisional clinical judgment, identifies missing or insufficiently supported evidence, asks targeted questions, and then regenerates the documentation from the augmented record. This Ask–then–Regenerate process is intended to better align with physicians' reasoning and help correct potential errors caused by missing evidence.

Each patient is documented by a clinical record X comprising an admission note and a progress note; we take the record as the model input X. Given a patient record X, we define four tasks in respiratory medicine.

Task A — Admission Diagnosis Generation. Generate structured admission diagnoses Y_A, with a primary diagnosis and optional secondary diagnoses.

Task B — Differential Diagnosis Discussion. Generate a free-text clinical analysis Y_B that integrates differential diagnosis, etiological reasoning, complication assessment, and the next-step plan.

Task C — Treatment Planning. Generate an itemized plan Y_C covering medications, examinations, nursing, and follow-up.

Task D — Active Clinical Questioning. Given the patient record X together with the model's preliminary outputs $(Y_A^{\text{init}}, Y_B^{\text{init}})$ from Tasks A and B, generate a set of questions $Q = \{q_1, \ldots, q_k\}$ targeting missing information from X or insufficient evidence in the preliminary differential diagnosis discussion.

These task dependencies define the two-stage pipeline in Figure 2. Stage I trains a unified Clinical Documentation Generator for Tasks A–C with interaction-guided multi-task SFT and an entity-weighted loss. Stage II trains the model for Task D and aligns its questioning behavior with physician preferences. At deployment, the system generates preliminary documentation, asks for missing evidence when needed, inserts the recovered answers into the corresponding record fields, and regenerates Tasks A–C from the augmented input. The closed-loop effect is evaluated with the Masked-Field Protocol by comparing complete, masked,

and question-filled records.

3.2 Stage I: Mitigating Task Discrepancies via Two-Stage Task-Interaction-Aware SFT

We characterize how Tasks A, B, and C interact through a small set of probe experiments, then use the resulting matrix to set the sampling proportions.

3.2.1 Quantifying Performance Discrepancies and Inter-Task Interference

We run controlled training experiments over Tasks A, B, and C under four configurations that share training steps, hyperparameters, and per-task example counts. In the **Single-task** configuration, each task is trained on its own data, providing a per-task reference. In the **Pairwise** configuration, the task pairs A+B, A+C, and B+C are each trained at a 1:1 sampling ratio. In the **Three-task** configuration, all three tasks are mixed at equal proportions to measure the aggregate effect. In the **Sequential** configuration, the tasks are trained one after another and all three are re-evaluated after each stage. We report the representative orderings A→B→C and C→B→A in the main text; the remaining four orderings showed the same qualitative pattern of sequential instability.

All configurations share training steps, hyperparameters, and per-task example counts. For pairwise comparisons, we quantify how task j affects task i on metric m as

$$T_{j\to i}^{(m)} = M_i^{(i,j)} - M_i^{(i)}, \tag{1}$$

where $M_i^{(i)}$ is the performance of task i trained alone, and $M_i^{(i,j)}$ is its performance when trained jointly with task j. A positive value indicates positive transfer; a negative value indicates interference. These values form the task interaction matrix $\mathbf{T}$.

We report ROUGE-L(Lin, 2004) and Entity Recall because they capture different evaluation aspects. ROUGE-L measures lexical and structural agreement with reference notes, whereas Entity Recall measures clinical content coverage. The clinical entities here are the disease, symptom, medication, laboratory-item, and procedure mentions annotated by respiratory medicine physicians in the reference notes; Entity Recall is the fraction of these reference entities recovered in the generated note, computed by the matching cascade detailed in Appendix C.1. We use Entity Recall as the main signal for mixing because missing clinical entities can change the clinical meaning of the generated documentation.

Task	BLEU-4	ROUGE-L	Entity Recall
A (Diagnosis)	0.16	0.24	71.64%
B (Diff. diagnosis)	0.03	0.16	87.46%
C (Treatment plan)	0.04	0.24	74.07%

Table 1: Single-task baseline performance.

Single-task references. Table 1 reports each task trained in isolation.

Asymmetric task interactions. Task B provides the strongest positive transfer: adding differential diagnosis discussion improves Task A Entity Recall from 71.64% to 79.85%, a gain of 8.21 points. This suggests that Task B data supplies disease and symptom evidence useful for diagnosis generation. The reverse is not symmetric: adding Task A lowers Task B's Entity Recall by 2.79 points. Task C is more stable as a target task, but adding it can reduce Task A ROUGE-L and Task B Entity Recall, likely because treatment plans use a more templated style and a different entity distribution.

Sequential instability. Sequential training is also unstable. In A→B→C, training on B first improves Task A, but the final C stage lowers Task B Entity Recall. In C→B→A, Task C similarly drops after the final A stage. We observed the same qualitative trend in the other task orderings, where the task trained in the final stage tends to suppress the Entity Recall of an earlier task. These results argue against naive sequential training and motivate a fixed, interaction-guided allocation for the final generator.

3.2.2 Task-Interaction-Aware Fine-tuning

We derive the sampling distribution from the positive-transfer entries of $\mathbf{T}$. The rule follows three constraints: (P1) every task keeps a baseline share; (P2) tasks that help other tasks receive extra weight; and (P3) negative transfers are handled implicitly by normalization rather than by explicit subtraction.

Let

$$\phi_i = \sum_{j\neq i} \max\left(T_{i\to j}^{(\text{Ent-R})}, 0\right) \tag{2}$$

measure the total positive transfer provided by task i. We assign each task a capped transfer reward,

$$w_i = 1 + \lambda\phi_i, \qquad p_i = \frac{w_i}{\sum_k w_k}. \tag{3}$$

The coefficient λ is set by a reward cap ρ:

$$\lambda = \frac{\rho - 1}{\max_i \phi_i}. \tag{4}$$

We set $\rho = 2$, so the strongest positive-transfer task receives at most twice the baseline weight. In our matrix, the dominant positive transfer is $T_{B\to A} = +8.21$, giving $\phi_B = 8.21$ and $\phi_A = \phi_C = 0$. The resulting weights are

$$w_A : w_B : w_C = 1 : 2 : 1, \tag{5}$$

and the final sampling distribution is

$$p_A : p_B : p_C = 1 : 2 : 1. \tag{6}$$

This converts the measured task interactions into a simple transfer-aware sampling prior for final multi-task fine-tuning.

3.2.3 Clinical Entity-Weighted Cross-Entropy Loss

Task mixing controls which examples the model sees. Entity-weighted loss controls which tokens matter more during learning. Standard cross-entropy treats all tokens equally, but clinical entities carry disproportionate weight: omitting a clinical entity can change the clinical meaning of a note.

We apply an entity-weighted focal loss:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T} \alpha_t \cdot (1 - p_t)^{\gamma} \cdot \log(p_t), \tag{7}$$

where p_t is the probability of the gold token, $\gamma = 2.0$, and $\alpha_t = 3.0$ if token t falls inside a clinical entity span and 1.0 otherwise. Entity spans are identified using the same clinical entity categories defined in Section 3.2.1 and aligned to subword tokens with the tokenizer offset map. Sample-level reweighting between physician-authored and augmented examples is detailed in Appendix B.3. This loss complements task-level mixing by emphasizing clinically meaningful tokens within each generated note.

3.3 Stage II: Proactive Clinical Questioning via Alignment and Feedback

Stage II builds proactive questioning in two phases: preference alignment for questioning and a closed-loop workflow that feeds recovered evidence back into documentation.

3.3.1 Clinical Preference Alignment for Questioning

Stage II trains Task D to ask questions that can improve downstream documentation. The model then targets the record and the preliminary admission diagnosis and differential diagnosis discussion outputs:

$$Q = \mathcal{M}_{\text{AQ}}(X, Y_A^{\text{pre}}, Y_B^{\text{pre}}). \tag{8}$$

The preliminary outputs expose the model's provisional assumptions. The questioning module then targets missing symptoms, risk factors, medication history, exposure history, prior treatment, tests, or other evidence that could affect the diagnosis or plan.

We align the questioning behavior with physician preferences. Ten respiratory medicine physicians annotate 200 de-identified cases. For each case, the SFT questioning model generates candidate questions conditioned on $(X, Y_A^{\text{pre}}, Y_B^{\text{pre}})$. Physicians label questions as clinically valuable or not based on clinical importance, specificity, and medical soundness. We form within-case preference pairs (q^+, q^-).

DPO(Rafailov et al., 2023) optimizes the questioning policy without a separate reward model:

$$\begin{aligned} \mathcal{L}_{\text{DPO}} &= -\,\mathbb{E}_{(x,q^+,q^-)}\Big[\log \sigma\big(\beta_{\text{DPO}}\, r\big)\Big], \\ r &= \log \frac{\pi_\theta(q^+ \mid x)}{\pi_{\text{ref}}(q^+ \mid x)} - \log \frac{\pi_\theta(q^- \mid x)}{\pi_{\text{ref}}(q^- \mid x)}, \end{aligned} \tag{9}$$

where $x = [X; Y_A^{\text{pre}}; Y_B^{\text{pre}}]$, π_θ is the optimized policy, π_{ref} is the frozen SFT reference model, and $\beta_{\text{DPO}} = 0.1$. The evaluation in Section 4.3 tests whether DPO improves the clinical value of the generated questions.

3.3.2 Ask–then–Regenerate with Masked-Field Evaluation

We integrate Task D into a closed-loop workflow targeting inputs that are under-documented at the source.

Step 1 — Initial generation. From the patient record X, the multi-task model produces preliminary outputs:

$$Y_A^{\text{draft}},\ Y_B^{\text{draft}},\ Y_C^{\text{draft}} = \mathcal{M}(X). \tag{10}$$

Step 2 — Questioning. The DPO-aligned model generates questions targeting gaps in X and weaknesses in the drafts:

$$Q = \mathcal{M}_{\text{DPO}}(X, Y_A^{\text{draft}}, Y_B^{\text{draft}}). \tag{11}$$

Step 3 — Augmentation. Answers $A(Q)$, supplied by the clinician at deployment or from a controlled source under evaluation, are inserted into the relevant record sections:

$$X' = X \oplus A(Q), \tag{12}$$

where $\oplus$ appends supplementary information to the relevant sections.

Step 4 — Regeneration. The same model regenerates documentation from the augmented record:

$$Y_A^{\text{final}},\ Y_B^{\text{final}},\ Y_C^{\text{final}} = \mathcal{M}(X'). \tag{13}$$

If Task D identifies clinically useful information gaps, regeneration from X' should improve over direct generation from X. We test this prediction in Section 4.4 with a Masked-Field Protocol and report gap closure:

$$\text{Gap Closure} = \frac{S(X^{\text{filled}}) - S(X^{\text{masked}})}{S(X^{\text{complete}}) - S(X^{\text{masked}})} \times 100\% \tag{14}$$

where S is a documentation score such as Entity Recall or ROUGE-L. The metric measures how much of the performance loss caused by missing input information is recovered under a fixed question budget.

4 Experiments

We evaluate the system following the two-stage structure of our method. We first assess the Stage I Clinical Documentation Generator on Tasks A–C (Section 4.2), then the Stage II active questioning model (Section 4.3), then the full Ask–then–Regenerate workflow under incomplete inputs (Section 4.4), and finally the contribution of each component (Section 4.5).

4.1 Experimental Setup

Base model. We use Qwen3-4B(Yang et al., 2025) as the base model. A respiratory medicine physician selected it from candidate open-source LLMs using an 80-question clinical knowledge evaluation (Appendix A.2). We fine-tune the model with LoRA (Hu et al., 2021), using rank $r = 64$, scaling factor $\alpha = 128$, dropout 0.05, and LoRA modules on all linear layers. Training uses 2 NVIDIA A6000 GPUs, per-device batch size 2, gradient accumulation 2, learning rate 1×10^{-4}, and cosine scheduling. Full hyperparameters are listed in Appendix A.1.

Data. We use 1000 de-identified respiratory-medicine inpatient records from a tertiary hospital in China. From these records we obtain 200 physician-annotated cases, split at the patient level into 100 for training and 100 for testing on Tasks A–C; the 100 training cases are expanded to 900 examples by an automated augmentation pipeline. For Task D, the same 200 records are additionally annotated by 10 respiratory medicine physicians with question-level preferences. Dataset construction, de-identification, the augmentation pipeline, and splits are detailed in Appendix B.

Metrics. For Tasks A–C, we report ROUGE-L, BLEU-4, and Entity Recall. Entity Recall is

$$\frac{|\mathcal{E}_{\text{gen}} \cap \mathcal{E}_{\text{ref}}|}{|\mathcal{E}_{\text{ref}}|}, \tag{15}$$

the fraction of reference clinical entities recovered in the generated note. Clinical entities include diseases, symptoms, medications, laboratory items, and procedures. For Task D, physicians rate clinical importance, specificity, and medical soundness on a 1–5 Likert scale. Under the Masked-Field Protocol, we also report Question Hit Rate and Key Fragment Recovery. Automatic metrics are averaged over 3 random seeds.

Baselines. Prompt baselines include Qwen3-4B zero-shot, GPT-4o and HuatuoGPT-II, all using the same task prompts. Training baselines use the same base model and data: separate single-task models, uniform multi-task mixing, sequential fine-tuning. Our method uses the Interaction-Guided Mixing rule from Section 3.2.

4.2 Interaction-Guided Multi-task Fine-tuning

We evaluate whether the Stage I Clinical Documentation Generator can perform the three documentation tasks in a unified multi-task setting. As shown in Table 2, Interaction-Guided Mixing improves average Entity Recall over uniform mixing (78.61 vs. 76.61) and avoids the forgetting observed in sequential training, where Differential Diagnosis Discussion drops to 66.07% under A→B→C and Treatment Planning drops to 67.52% under C→B→A. Our method keeps all three tasks close to or above their separate single-task references within one unified generator. Compared with zero-shot Qwen3-4B, it improves average Entity Recall by 9.50 points, and also outperforms GPT-4o on

Strategy	Diag R-L	Diag Ent-R	Reason R-L	Reason Ent-R	Plan R-L	Plan Ent-R
Separate single-task models	0.24	71.64	0.16	87.46	0.24	74.07
Qwen3-4B zero-shot	0.18	70.51	0.12	69.05	0.19	67.77
GPT-4o	0.20	72.30	0.14	74.50	0.21	70.20
HuatuoGPT-II	0.17	68.90	0.13	71.20	0.18	66.50
Uniform mixing	0.18	75.36	0.18	82.32	0.23	72.14
Sequential A→B→C	0.12	73.88	0.14	66.07	0.26	72.21
Sequential C→B→A	0.18	70.90	0.17	71.72	0.20	67.52
Ours	**0.23**	**77.45**	**0.17**	**84.26**	**0.24**	**74.13**

Table 2: Results on the three documentation tasks. Ent-R: Entity Recall (%). R-L: ROUGE-L. Separate single-task models use three independent models and are references rather than deployable unified systems. Fine-tuned results are averaged over 3 seeds.

average Entity Recall (78.61 vs. 72.33), supporting the effectiveness of Stage I training.

4.3 Physician-Annotated Evaluation of Active Questioning

We evaluate active questioning using physician ratings and masking-log metrics. Physicians rate Clinical Importance, Specificity, and Medical Soundness to assess whether each question is necessary, targeted, and medically valid; Question Hit Rate and Key Fragment Recovery measure whether generated questions match and recover deleted fragments.

As shown in Table 3, DPO improves both physician-rated question quality and evidence recovery. Compared with SFT only, SFT+DPO raises the average physician rating from 4.21 to 4.26 and increases Key Fragment Recovery from 44% to 56%, suggesting that preference alignment helps the model ask questions that recover clinically useful missing evidence.

4.4 Closed-Loop Evaluation under Incomplete Records

We evaluate the full closed-loop workflow with the Masked-Field Protocol (Appendix C.2). We delete clinically important fragments from complete held-out records X^{complete} to produce X^{masked}, with deletion done by physical removal of the text span rather than paraphrasing. The oracle is question-conditioned: a removed fragment is returned only when a generated question matches it in the masking log, and matched fragments are reinserted into their original record fields to form X^{filled}. This design prevents the oracle from revealing all masked information and tests whether the model asks for the evidence that matters.We compare three conditions: direct generation from X^{masked}; Ask–then–Regenerate from X^{filled}; and generation from X^{complete} as an upper bound.

Ask–then–Regenerate improves Entity Recall on all three tasks over masked-input generation, with the largest gains on diagnosis and clinical reasoning (Table 4). It closes 52.71% of the average Entity Recall gap to complete-input generation, showing that the model does not merely ask plausible questions but asks for information that improves downstream documentation.To make the closed-loop evaluation more interpretable, Appendix D provides a qualitative case study, where Figure 3 illustrates the masking and questioning process, and Figure 4 shows how the recovered evidence is used to revise the final diagnosis.

4.5 Ablation and Analysis

We ablate the components that support the generator, active questioning, and the closed-loop workflow. Entity Recall is measured on the full test set; Gap Closure is measured on the 10-case Masked-Field held-out set.

Removing Ask–then–Regenerate causes the largest drop under incomplete inputs (Table 4), confirming that information recovery drives the main result. Removing DPO reduces gap closure by 13.71 points (Table 5) and lowers the average physician rating by 0.05 points (Table 3), showing that better questions translate into better regenerated notes. Replacing Interaction-Guided Mixing with uniform mixing reduces full-test-set average Entity Recall by 2.00 points, consistent with the task interaction analysis in Section 3.2. Removing entity-weighted loss leaves ROUGE-L nearly unchanged but lowers full-test-set Entity Recall by 1.28 points.

Model	Physician-Rated Clinical Evaluation				Masked-Field Automatic Evaluation	
	Physician Avg.	Clinical Importance	Specificity	Medical Soundness	Question Hit Rate	Key Fragment Recovery
Qwen3-4B zero-shot	4.15	3.32	4.62	4.50	28%	35%
SFT only	4.21	3.38	4.72	4.52	34%	44%
GPT-4o	4.21	3.40	4.68	4.56	33%	42%
SFT + DPO (Ours)	**4.26**	**3.44**	**4.80**	**4.54**	**41%**	**56%**

Table 3: Active questioning evaluation. Physician-rated clinical evaluation reports 1–5 Likert scores annotated by respiratory medicine physicians. Masked-field automatic evaluation reports recovery metrics computed from the masking log.

Condition	Diag Ent-Recall	Reason Ent-Recall	Plan Ent-Recall	Avg Ent-Recall
Masked input	73.33	70.51	72.20	72.01
Ask–then–Regenerate	80.00	77.42	72.92	76.78
Complete input	86.67	82.49	74.01	81.06
Gap Closure	50.00%	57.68%	39.78%	52.71%

Table 4: Closed-loop evaluation under the Masked-Field Protocol. Gap Closure is computed separately for each task using Entity Recall across the masked, ask–then–regenerate, and complete-input conditions.

5 Conclusion

We presented Ask–then–Regenerate, a two-stage framework for clinical documentation under incomplete admission records. It first trains a unified generator for three documentation tasks, then trains active questioning to recover missing evidence before regeneration. Experiments show that interaction-guided fine-tuning improves entity coverage, and the Masked-Field Protocol shows that active questioning recovers 52.71% of the average Entity Recall loss. These results suggest that asking before writing improves robustness over single-pass generation.

Limitations

Several limitations should be noted. First, our system is developed and evaluated on respiratory medicine data from a single tertiary hospital, so its gains may not fully generalize to other specialties, institutions, or larger base models. Second, the Masked-Field Protocol uses an LLM judge for segmentation and question matching; although the procedure was reviewed by a respiratory medicine physician, residual judge bias cannot be fully ruled out. Third, the current workflow is single-turn and manually answered; multi-turn, EHR-integrated questioning remains future work.

References

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Huatuogpt-ii, one-stage training for medical adaption of llms. *Preprint*, arXiv:2311.09774.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, huimin zheng, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, and Xiangmin Xu. 2023. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *Preprint*, arXiv:2310.15896.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Yusheng Liao, Shuyang Jiang, Zhe Chen, Yu Wang, and Yanfeng Wang. 2024. MedCare: Advancing medical LLMs through decoupling clinical alignment and

Variant	Diag Ent-Recall	Reason Ent-Recall	Plan Ent-Recall	Avg Ent-Recall	Avg Ent-Recall Gap Closure
Full system	77.45	84.26	74.13	78.61	52.71%
w/o entity-weighted loss	77.60	82.80	71.60	77.33	47.82%
w/o interaction-guided mixing	75.36	82.32	72.14	76.61	50.33%
w/o DPO	75.10	78.30	72.40	75.27	39.00%

Table 5: Ablation study. The three Entity Recall columns are measured on the full test set; the Gap Closure column is measured on the 10-case Masked-Field held-out set, where Gap Closure is defined. The masked-input condition is reported in Table 4.

knowledge aggregation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10562–10581, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. 2025. Regmix: Data mixture as regression for language model pre-training. In *International Conference on Learning Representations*, volume 2025, pages 38305–38339.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. Meddg: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. *Preprint*, arXiv:2010.07497.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia. Association for Computational Linguistics.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2023. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36:69798–69818.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. *Preprint*, arXiv:2109.12174.

A Training Details

A.1 Training Configuration and Hyperparameters

We fine-tune the model in two stages. Stage I performs multi-task supervised fine-tuning for Tasks A–C; Stage II performs DPO for active questioning on top of the Stage I checkpoint. Both stages use LoRA and bf16 mixed precision, with gradient checkpointing enabled to reduce memory. Stage I results are averaged over three random seeds (42, 123, 456). All hyperparameters are listed in Table 6.

Hyperparameter	Stage I (SFT)	Stage II (DPO)
Base model	Qwen3-4B	Stage I model
Fine-tuning	LoRA	LoRA
LoRA rank r	64	16
LoRA α	128	32
LoRA dropout	0.05	0
LoRA target	all linear	all linear
Learning rate	1×10^{-4}	5×10^{-6}
LR scheduler	cosine	cosine
Warmup ratio	0.1	0.1
Epochs	3	3
Per-device batch size	2	1
Gradient accumulation	2	16
Max sequence length	4096	4096
Optimizer	AdamW	AdamW
Precision	bf16	bf16
Preference β	—	0.1
Preference loss	—	sigmoid (DPO)

Table 6: Training hyperparameters for the two-stage pipeline.

Candidate Model	Params	Answer Rate (%)
Fleming-R1-7B	7B	20.0
Qwen2.5-7B-Instruct	7B	100.0
Qwen3-8B	8B	92.5
Qwen3-4B-Thinking-2507	4B	93.8
Qwen3-4B-Instruct-2507 (ours)	4B	**100.0**

Table 7: Base model selection on the clinical knowledge probe. Answer Rate is the fraction of the 80 concepts the model explained instead of abstaining with "I don't know."

A.2 Base Model Selection

We selected the base model with a clinical knowledge probe of 80 short-answer questions on respiratory medicine concepts, covering diseases, examinations, and related clinical terms. Each question takes the form "What is X? Answer 'I don't know' if you are not sure.", so that a model unable to explain a concept abstains rather than fabricates an answer. A respiratory medicine physician reviewed the responses of five candidate open-source LLMs and assessed their clinical quality. Table 7 reports the fraction of concepts each model answered instead of abstaining. Fleming-R1-7B abstained on most questions, indicating weak coverage of the target domain, whereas the Qwen-family models covered nearly all concepts. Among the strong candidates, the physician judged Qwen3-4B-Instruct-2507 to give the most clinically appropriate answers; we adopt it (referred to as Qwen3-4B in the main text), as it also has the smallest parameter count, which is preferable for hospital-side deployment.

B Dataset Construction

B.1 Data Source and De-identification

Our data consists of de-identified inpatient records collected from the respiratory medicine department of a tertiary hospital in China between 2020 and 2024.[2] Each patient has two documents: an admission note and a progress note. The admission note contains the chief complaint, history of present illness, past medical history, personal history, physical examination, and auxiliary tests, along with the attending physician's admission diagnosis. The progress note comprises the first progress note and the subsequent daily progress notes recording the patient's condition during hospitalization; the reference differential diagnosis discussion and treatment plan for Tasks B and C are extracted from the first progress note. All references were verified by respiratory medicine physicians. All records were de-identified before use: personal identifiers such as patient names, ID numbers, and other private fields were removed by regular expression matching. The use of these data was approved by the ethics committee of the participating hospital, with informed consent waived for the retrospective use of de-identified records. Due to patient privacy and hospital policy, the dataset cannot be publicly released.

B.2 Data Augmentation Pipeline

Because physician-annotated examples are limited, we build an automated pipeline to expand the training set. Given a de-identified record, we prompt Qwen3-30B-A3B-Instruct-2507 with few-shot examples to generate the task outputs; for each task we produce both a concise, physician-style version and a detailed, point-by-point version, and the final generator is trained on both. The generated drafts then pass a multi-stage quality control procedure consisting of rule-based cleaning (removing empty, garbled, and duplicated text and re-applying de-identification), per-task length constraints, repetition detection, and a semantic consistency check that retains a sample only when the SBERT cosine similarity between its concise and detailed versions is at least 0.85. Samples that fail any stage are routed to a retry step for regeneration and re-auditing.

[2] The hospital name is withheld to preserve anonymity during review and will be disclosed in the camera-ready version.

B.3 Dataset Statistics and Splits

We split the 200 physician-annotated cases at the patient level into two disjoint halves: 100 cases form the basis of the training set, and the remaining 100 cases are held out as the test set for Tasks A–C. The 100 training cases are expanded by the augmentation pipeline into 800 additional examples; the training set therefore contains 900 examples in total, of which 10% are held out as a validation set for checkpoint selection. Since the split is performed at the patient level, no patient appears in both the training and test sets. For the active questioning task (Task D), the same 200 records are additionally annotated by 10 respiratory medicine physicians with question-level preference pairs; this preference annotation forms a separate layer on top of the records, and Task D is evaluated on its own split rather than on the 100-case test set used for Tasks A–C. Training examples are weighted at the sample level to balance physician-authored and augmented data.

C Evaluation Protocol

C.1 Entity Recall Computation

Clinical entities in the reference notes are annotated by physicians and cover diseases, symptoms, medications, laboratory items, and procedures. For each generated note, every reference entity is matched against the generated text through a cascade of four matchers, tried in order: exact string match; synonym match against a curated medical synonym dictionary (e.g., abbreviations and common variants); partial match, which fires when at least 70% of an entity's sub-tokens appear in the text; and semantic match, where the entity and candidate text spans are encoded with a Chinese medical embedding model and matched when their cosine similarity reaches 0.85. Before matching, all text is converted to simplified Chinese and whitespace-normalized. An entity is counted as recalled if any matcher succeeds, and Entity Recall is the fraction of reference entities recalled. This recall-oriented metric reflects whether clinically important entities are preserved in the generated documentation.

C.2 Masked-Field Protocol

The Masked-Field Protocol constructs controlled information gaps in held-out records and is implemented as two pipelines.

Masking. Masking is applied to four fields of the admission record: history of present illness, past medical history, personal history, and admission condition. Each field is segmented into fragments by an LLM, which labels every fragment with one of five types: diagnostic clue, treatment history, risk factor, context, or narrative. Each fragment is then deleted with probability $p_{\text{delete}} = s \cdot w_{\text{type}}$, where s is a global mask strength and w_{type} is a type-dependent weight: diagnostic clue 1.0, treatment history 0.7, risk factor 0.5, context 0.2, and narrative 0.0. Narrative fragments, which carry the temporal backbone of the record, are therefore never removed. Deletion is physical removal of the text span rather than paraphrasing. We sweep s over 0.2–1.0 and use $s = 0.4$ for the main experiments. A masking log records all fragments, including their original text and position, enabling exact-position restoration during fill-in.

Question-conditioned fill-in. Given the questions generated for a masked record, each question is matched by an LLM against the set of deleted fragments; the LLM either selects one fragment or returns `NO_MATCH`. A matched fragment is reinserted at its original position in the corresponding field, and each fragment is restored at most once even if hit by multiple questions. Unmatched fragments are not returned, so the oracle reveals only the information the model explicitly asks for.

Judge model and validation. Both the fragment segmentation and the question-fragment matching are performed by Qwen3-30B-A3B-Instruct-2507. The full procedure and its intermediate outputs were reviewed by a respiratory medicine physician, who confirmed that the segmentation and matching results were clinically reasonable. We did not conduct a quantitative inter-annotator agreement study, so residual judge bias cannot be fully ruled out.

C.3 Question-Level Metrics

Under the Masked-Field Protocol we report two automatic metrics derived from the masking log. Question Hit Rate is the fraction of generated questions that match at least one deleted fragment. Key Fragment Recovery is the fraction of deleted fragments that are recovered by the generated questions. The closed-loop evaluation in the main text is conducted on 10 held-out cases.

C.4 Physician Rating

For the active questioning task, physicians rate each generated question on a 1–5 Likert scale along the three dimensions defined in the scoring rubric (Appendix E): clinical importance, specificity, and medical soundness.

D Case Study of Ask–then–Regenerate

Appendix D presents a qualitative example of the Ask–then–Regenerate workflow under the Masked-Field Protocol. The case involves a 67-year-old male patient with chronic cough and sputum production for more than two years. In the complete record, the patient had repeated CT evidence of bilateral multilobar bronchiectasis with infection, elevated airway inflammation markers, negative pathogen and TB/NTM-related tests, progressive respiratory symptoms, and limited response to prior anti-infective treatment. During masking, several of these clinically important fragments were removed, leaving only partial evidence such as obstructive ventilatory dysfunction, elevated T-IgE, sinusitis, inflammatory markers, poor treatment response, and weight loss. As a result, the initial generation from the masked record overemphasized COPD as the primary diagnosis. The system then asked targeted questions about chest imaging, hemoptysis, tuberculosis activity, dyspnea burden, and microbiological testing. After the recovered evidence was inserted back into the record, the regenerated diagnosis shifted toward bronchiectasis with infection, which was closer to the reference diagnosis. This example illustrates how active questioning can help recover missing evidence and correct the diagnostic focus before final documentation is produced.

E Prompts

This section lists the prompts used for the four tasks and the rubric used by physicians to score candidate questions for DPO. The prompts are translated from the Chinese originals used in our experiments, in which the clinical records and the model operate in Chinese. The same task prompts are used for our model and for the prompt-based baselines.

E.1 Prompt for Admission Diagnosis Generation (Task A)

You are a clinical decision-support assistant. Based on the patient's admission information, generate the admission diagnoses. The output must contain one primary diagnosis and, if applicable, several secondary diagnoses.
- The primary diagnosis should be the most central and clearly established disease.
- If the primary diagnosis is clear, first list complications or differential diagnoses directly related to it as secondary diagnoses, then list any other conditions the patient has.
- If a secondary diagnosis is subsumed by the primary diagnosis, omit it from the secondary list to keep the output concise and accurate.

Admission information: {ADMISSION_RECORD}

Complete admission record

67-year-old male with cough and sputum for >2 years.

Present illness:
Chronic cough with white sticky sputum since 2018, occasionally accompanied by shortness of breath and throat itching. Symptoms could be triggered by irritant odors. In 2020, cough worsened with difficult sputum expectoration, nocturnal awakening, sweating during attacks, and 10-kg weight loss over 2 years.

Available evidence:
- Pulmonary function: obstructive ventilatory dysfunction; FEV1 2.79 L, FEV1/FVC 67.8%; bronchodilator test negative.
- Imaging: repeated chest CT showed bilateral multilobar bronchiectasis with infection, chronic bronchitis, emphysema, and multiple small nodules in the left upper lung.
- Airway inflammation / allergy markers: FeNO 168 ppb; T-IgE elevated (up to 847 KU/L).
- Previous evaluation: sinus CT showed multiple sinusitis; superficial lymph node ultrasound found a right supraclavicular hypoechoic nodule.
- Microbiology: routine inflammatory markers and sputum/pathogen tests, including bacterial/fungal smear and culture, acid-fast smear, and TB/NTM-DNA, showed no clear abnormality.
- Prior treatment response: limited improvement after levofloxacin and later itraconazole.

Relevant history:
Old pulmonary tuberculosis, hypertension, allergic rhinitis, smoking history (30 years, quit 10 years ago), alcohol use, cat exposure, and occupational exposure to pesticides and boiler work.

Other comorbidities:
Chronic gastritis, gastroesophageal reflux disease, carotid atherosclerosis, conductive hearing loss, and prior thyroid/parotid surgery.

Masked admission record

67-year-old male with cough and sputum for >2 years.

Available evidence:
FEV1/FVC 67.8%, CAT 13, elevated T-IgE, sinusitis, ANA(+), ESR/LDH elevation, poor response to prior treatment, and 10-kg weight loss.

[MASKED]:
- chest CT evidence of bilateral multilobar bronchiectasis with infection;
- FeNO 168 ppb and peak T-IgE 847 KU/L;
- repeated imaging showing bronchiectasis involving multiple lobes;
- sputum/pathogen tests and negative TB/NTM-related evidence;
- symptom progression with difficult sputum expectoration and nocturnal awakening.

Questions asked by the active module

- For the two-year history of cough and sputum, did the patient have recurrent hemoptysis or blood-streaked sputum? If so, what were the frequency, amount, color, and triggers?
- During the 2019 hospitalization, were chest imaging examinations such as chest CT or HRCT performed? Did they show bronchiectasis, pulmonary infection, emphysema, or pulmonary nodules?
- Given the history of old pulmonary tuberculosis, did the patient have fever, night sweats, weight loss, or any microbiological or imaging evidence suggesting active tuberculosis?
- Beyond the CAT score of 13, what was the patient's actual dyspnea burden and activity tolerance? Were there symptoms such as nocturnal awakening, exertional worsening, or objective pulmonary function changes?
- Since chronic fungal infection was suspected, were sputum microbiology, fungal smear/culture, fungal antigen testing, acid-fast smear, or TB/NTM tests performed? What were the results?

Figure 3: A case study of the Masked-Field Protocol. The complete admission record contains repeated CT evidence of bronchiectasis with infection, airway inflammation markers, microbiological tests, symptom progression, and relevant exposure history. The masked record removes several clinically important fragments, after which the active questioning module asks targeted questions to recover the missing evidence.

Question-filled record

Recovered evidence inserted back into the record:
- repeated chest CT showed bilateral multilobar bronchiectasis with infection;
- FeNO 168 ppb and T-IgE up to 847 KU/L;
- sputum/pathogen tests, including bacterial/fungal studies and TB/NTM-related tests, showed no clear abnormality;
- symptoms progressed to difficult sputum expectoration and nocturnal awakening;
- prior anti-infective and antifungal treatment resulted in limited symptom relief.

Reference diagnosis

Primary diagnosis: Bronchiectasis with infection

Secondary diagnoses:
COPD, hypertension, carotid atherosclerosis, allergic rhinitis, sinusitis, conductive hearin

Initial diagnosis from masked record

Primary diagnosis: Chronic obstructive pulmonary disease (COPD)

Rationale:
- The patient had chronic cough and sputum for >2 years, with a smoking history and obstructive ventilatory dysfunction (FEV1/FVC 67.8%).
- The available evidence supported chronic airway disease, but key imaging and infection-related evidence for bronchiectasis was missing.

Secondary diagnoses:
Bronchiectasis, chronic sinusitis, allergic rhinitis, hypertension, old pulmonary tuberculosis, ...

Regenerated diagnosis from question-filled record:

Primary diagnosis: Bronchiectasis with infection

Rationale:
- The recovered CT findings showed bilateral multilobar bronchiectasis with infection.
- Together with chronic cough/sputum, progressive symptoms, and poor response to prior treatment, the evidence more strongly supported bronchiectasis-related infection than COPD alone.

Secondary diagnoses:
COPD, allergic rhinitis, sinusitis, hypertension, carotid atherosclerosis, ...

Figure 4: Diagnosis revision under Ask–then–Regenerate. From the masked record, the model initially over-prioritizes COPD as the primary diagnosis because key bronchiectasis-related evidence is unavailable. After the recovered evidence is inserted back into the record, regeneration shifts the primary diagnosis toward bronchiectasis with infection, closer to the reference diagnosis.

E.2 Prompt for Differential Diagnosis Discussion (Task B)

You are a professional AI medical assistant. Based on the patient's initial admission information, generate a clinical reasoning discussion covering three parts: diagnostic evidence, etiological analysis, and differential diagnosis.

First assess how well the diagnosis is established:
- Established: supported by definitive findings (e.g., pathological confirmation or a positive specific test).
- Not established: only symptom descriptions, without specific diagnostic evidence.

Adjust the focus of the discussion according to the level of certainty:
- Unclear before admission, established after admission: emphasize the key confirmatory findings and the diagnostic process; analyze the cause of the final diagnosis; briefly review the differential considered at admission.
- Diagnosis established at admission, urgent treatment needed: state the established evidence concisely; analyze the disease mechanism; focus on the differentiation of complications and treatment priority.
- Establishing the primary diagnosis while managing complications: address evidence for both the primary diagnosis and the complications; analyze their link; balance confirmation of the primary diagnosis with differentiation of complications.

Admission information: {ADMISSION_RECORD}

E.3 Prompt for Treatment Planning (Task C)

You are a professional clinical decision-support assistant. Based on the patient's current clinical situation, generate a treatment plan.

The output must contain exactly two parts, each beginning with the marker "[Detailed Version]" or "[Concise Version]", which is essential for downstream processing.
- Detailed version (structured): comprehensively cover examinations, treatment (pharmacological / non-pharmacological / surgical), nursing and monitoring, special management, and temporal priority; state the supporting evidence for each item.
- Concise version (key points): keep only the highest-priority and clinically essential actions; drug names, dosages, and core examination items must be consistent with the detailed version; use a numbered list and remove redundant explanation.

Output template:

```
[Detailed Version]
 Treatment Plan (structured)
1. Examination plan
2. Treatment (pharmacological /
non-pharmacological / surgical)
3. Nursing and monitoring
4. Special communication
5. Treatment sequence and priority

[Concise Version]
(numbered list, key actions only)
```

Clinical situation: {ADMISSION_RECORD}

E.4 Prompt for Active Questioning (Task D)

You are an experienced clinician conducting an in-depth review of an admission record. Together with the admission diagnosis, check whether the diagnostic evidence is sufficient and whether key information is missing, and ask follow-up questions to recover the missing information.

Task: based on the patient record below, generate exactly 5 key questions, used to verify the soundness of the admission diagnosis, supplement essential diagnostic evidence, and clarify contradictions between the record and the diagnosis.
- If the current diagnosis holds: check whether the key symptom features, triggers, progression patterns, and specific risk factors supporting the diagnosis are sufficiently documented.
- If the current diagnosis is questionable: ask for the key information needed to refine the differential diagnosis and to support or rule out the diagnosis.
- Do not fabricate information not present in the record, and do not repeat questions about content already clearly documented.

Admission record: {ADMISSION_RECORD}
Primary diagnosis: {MAIN_DIAGNOSIS}
Secondary diagnosis: {SECONDARY_DIAGNOSIS}

Output standard JSON only, with exactly 5 questions, in the following structure:

```
{
  "questions": ["question1",
                "question2",
                "question3",
                "question4",
                "question5"]
}
```

E.5 Physician Scoring Rubric for DPO

Physicians rate each candidate question on a 1–5 scale along the following three dimensions; the ratings are used to construct DPO preference pairs.

Clinical Importance — Should this question be asked, and how much would the answer affect the clinical decision?
5 = Must ask (directly affects diagnosis or treatment); 4 = Should ask (significantly improves the decision); 3 = May ask (useful but not critical); 2 = Need not ask (limited impact); 1 = Should not ask (irrelevant or already recorded).

Specificity — Is the question precise and targeted?
5 = Highly specific (points to a specific symptom / finding / time point); 4 = Fairly specific; 3 = Moderate (clear direction, lacking qualifiers); 2 = Broad (applies to any record); 1 = Completely vague.

Medical Soundness — Is the question medically correct in terminology and logic?
5 = Fully sound; 4 = Mostly sound (minor imprecision); 3 = Moderate (small flaws); 2 = Problematic (incorrect terminology or logic); 1 = Seriously flawed and potentially misleading.