Aligning Small Language Models for Clinical Note Generation via Minimal Edit Preference Optimization

Anonymous ACL submission

Abstract

Automatic clinical note generation leveraging Large Language Models (LLMs) significantly alleviates the administrative burden on clinicians. However, adopting proprietary LLMs poses substantial risks to sensitive patient information, while deploying large-scale open-source models remains impractical due to prohibitive costs. Although Small Language Models (SLMs) offer a cost-effective alternative for on-device implementation, a formidable performance gap persists compared to their larger counterparts. Knowledge distillation is widely used to bridge this gap, yet conventional methods such as supervised fine-tuning often fail to capture the intricate nuances of the clinical domain. In this work, we propose Minimal Edit Preference Optimization (**MEPO**), a novel distillation framework optimized for clinical note generation. By focusing on generating fine-grained preference signals to address common SLM errors in clinical note generation, we achieve a 4.5% relative improvement in the key metric over the DPO baseline, demonstrating that SLMs can achieve high fidelity in specialized medical documentation.

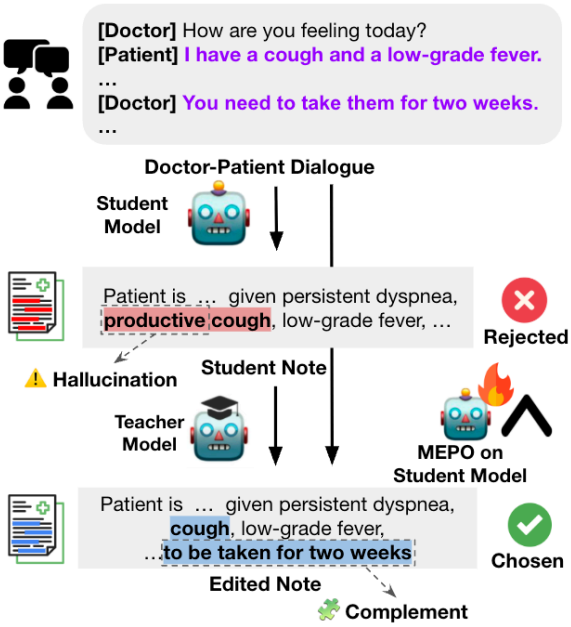

Figure 1: Overview of our framework. The teacher model refines student notes by rectifying hallucinations and adding missing information. The resulting teacher-edited note ('chosen') and the original student note ('rejected') are used for preference optimization.

1 Introduction

Clinical note generation is one of the most promising medical applications of Large Language Models (LLMs), as it can reduce documentation burden and allow clinicians to spend more time on direct patient care (Finley et al., 2018; Enarvi et al., 2020; Molenaar et al., 2020; Ahn et al., 2024). Recent proprietary LLMs like ChatGPT exhibit strong performance (Giorgi et al., 2023; Yim et al., 2023; Bedi et al., 2025), but their use in healthcare is constrained by privacy, compliance, and data-governance concerns. Open-source alternatives are therefore attractive, demonstrating performance comparable to proprietary models; however, they face practical limitations due to the high memory requirements for inference. To address these hardware limitations, Small Language Models (SLMs) have emerged as a viable solution for privacy-preserving, on-premise deployment. However, SLMs still underperform larger models in clinical reasoning quality.

Knowledge distillation (Hinton et al., 2015) is an effective strategy to close this gap, but standard pipelines-especially those involving RLHF-require prohibitive human annotation or complex multi-stage training (Ouyang et al., 2022; Bai et al., 2022). Due to the scalability limitations of manual preference labeling, prior work has explored alternatives such as DistillDirect (Wang et al., 2025a), which leverages output from a stronger teacher model as a preference signal.

However, recent analysis has shown that imitation-based distillation tends to replicate the teacher's surface-level style rather than its factual competence, leaving a substantial capability gap that cannot be bridged simply by scaling the amount of imitation data (Gudibande et al., 2023). This limitation is particularly concerning in clinical settings where stylistic fluency can mask factual errors in medical reasoning. Moreover, naively treating teacher output as a chosen response can distort the model's probability distribution and lead to suboptimal alignment (Wang et al., 2025b). Instead, leveraging edits can be effective since minimal, targeted corrections preserve the original response distribution while providing actionable supervision, and post-edits can be repurposed as implicit preference signals that better capture fine-grained revision preferences (Berger et al., 2024; Liang et al., 2025).

In this paper, we propose **MEPO**, an efficient distillation framework for clinical note generation that combines teacher-guided preference labeling with minimal-edit supervision. By generating student notes once and applying targeted corrections, MEPO improves training efficiency and domain fidelity without repeated on-policy sampling. Our experimental results demonstrate that MEPO achieves a 4.5% relative improvement in average clinical utility scores over the standard DPO baseline, closing the gap between SLMs and large-scale models in medical documentation.

2 Methodology

2.1 Problem Formulation

We formulate the task as generating the Assessment and Plan (AP) section of a SOAP (Subjective, Objective, Assessment, Plan) note from a physician-patient dialogue transcript. Prior work (Wang et al., 2025a) has demonstrated that generating a Subjective section is no longer challenging enough to justify a specialized model. For the Objective section, most data are integrated into Electronic Medical Records (EMR) and are easily incorporated into a note without a specialized LLM. Therefore, we focus on AP section generation due to its clinical complexity. The Assessment section summarizes and interprets symptoms, problem interactions, and longitudinal context across encounters, while the Plan section includes medications, tests, referrals, and follow-up instructions, requiring correct clinical intent and dosage specification. Because of the representational gap between patient-centered dialogue and clinician-oriented documentation, generating the AP section involves transforming information across communicative registers, which makes it a challenging task for SLMs (Stupp et al., 2022).

2.2 Minimal Edit Preference Optimization

We observe that SLMs frequently produce plausible notes that nevertheless contain factual inaccuracies, unsupported elaborations, or omissions of clinically important details. To address such failures under the lack of reliable open resources tailored for clinical note generation, we propose Minimal Edit Preference Optimization (MEPO). MEPO is a novel distillation framework, consisting of 1) edit-based data synthesis for obtaining robust training signals, and 2) Direct Preference Optimization (DPO) training. Our proposed method for constructing a preference dataset is as follows.

For each dialogue x, we first sample an AP-section focused clinical note from the teacher model T and the trainable student model π_θ respectively. The teacher-generated note $\hat{y}$ serves as a reference during the minimal edit process. The student-generated note y^- exposes diverse note-generation errors and serves as the sub-optimal note onto which edits are applied. The prompt we use to generate clinical notes is shown in Table 3.

Each student note y^- gets paired with the corresponding dialogue transcript x and is provided to the teacher model T acting as a clinical editor. To facilitate logical corrections, we also provide the corresponding teacher note $\hat{y}$ to the T. During the minimal editing process, the teacher model T is instructed to 1) evaluate the student note y^- against the dialogue x through explicit reasoning, 2) identify erroneous sections and present evidence in the dialogue, 3) conduct minimal edits to fix any factual inaccuracies, medical hallucinations, and clinically important omissions, while preserving the student note's tone and structure as much as possible, and 4) generate the final edited note y^+.

In order to guide the teacher model's editing process, we constrain each editing process to take place within pre-defined XML tags such as `<reasoning>`, `<modified>`, `<evidence>`, and `<note>`. `<reasoning>` is placed up front to elicit high-level evaluation of the note, and actual edits that follow are enclosed in `<modified>` for explicit separation. `<evidence>` is required for every edit, with the intent to ground each edit on

Method / Setting	Validation Set				Test Set			
	Acc	Comp	Cla	Avg	Acc	Comp	Cla	Avg
Teacher Model								
Qwen3-Next-80B-A3B-Instruct	4.72	4.30	4.82	4.61	4.71	4.15	4.73	4.53
Student Models								
Qwen2.5-7B-Instruct	4.27	3.27	4.10	3.88	4.23	3.26	4.12	3.87
w/ SFT	4.37 +0.10	4.08 +0.81	4.52 +0.42	4.32 +0.44	4.38 +0.15	3.87 +0.61	4.52 +0.40	4.26 +0.39
w/ DPO	4.30 +0.03	3.45 +0.18	4.20 +0.10	3.98 +0.10	4.23 +0.00	3.42 +0.16	4.15 +0.03	3.93 +0.06
w/ SFT + DPO	4.30 +0.03	4.12 +0.85	4.55 +0.45	4.32 +0.44	4.16 -0.07	3.94 +0.68	4.46 +0.34	4.19 +0.32
w/ MEPO (Ours)	4.35 +0.08	3.90 +0.63	4.42 +0.32	4.22 +0.34	4.21 -0.02	3.86 +0.60	4.33 +0.21	4.13 +0.26
w/ SFT + MEPO (Ours)	**4.53** +0.26	**4.18** +0.91	**4.58** +0.48	**4.43** +0.55	**4.53** +0.30	**4.04** +0.78	**4.56** +0.44	**4.38** +0.51

Table 1: Main experimental results. Scores range from 1 to 5. Subscripts (tiny text) indicate the performance margin compared to the base Student model (Qwen2.5-7B). Our MEPO-based framework significantly outperforms conventional distillation baselines.

the dialogue. To encourage localized corrections and avoid excessive rewriting, the teacher is instructed to present both the target span and the edited span, enclosed in `<before>` and `<after>`, for each edit. We parse `<note>` section from raw output as edited note y^+ and construct DPO pair as: $y^+ =$ edited note, $y^- =$ original policy note. We name the minimal-edit-based preference labeling and subsequent DPO training framework as Minimal Edit Preference Optimization (MEPO). The prompt for minimal edit is shown in Table 9, and we present examples through Table 12 and Table 15. A detailed analysis of our dataset can be found in Appendix E.

3 Experimental Setup

3.1 Models and Baselines

We use Qwen3-Next-80B-A3B-Instruct-FP8 (Qwen Team, 2025) as teacher T for minimal editing and teacher note generation, and Qwen2.5-7B-Instruct (Yang et al., 2024) as student. We compare our approach against the standard Supervised Fine-Tuning (SFT) approach and the DPO approach from previous work (Wang et al., 2025a). We list detailed training settings in Appendix A

3.2 Dataset

Data Splits. Following LLaMA-Clinic (Wang et al., 2025a), we source our dataset from ACI-Bench and Dialog-G. ACI-Bench is a clinical dialogue summarization benchmark containing real physician-patient encounters paired with expert-written notes (Yim et al., 2023), and Dialog-G is a dialogue-note dataset generated from MT-Samples clinical materials (Wang et al., 2025a). We use 1,358 unique dialogues from the ACI-Bench train set (67) and Dialog-G (1,291) to generate our own targets and preference pairs for training. For validation and test, we use the ACI-Bench valid set (20) and the ACI-Bench test sets (120).

Data construction. For SFT dataset construction, T generates a teacher note $\hat{y}$ for each x, yielding a supervision dataset in the form of $(x, \hat{y})$. Examples are shown in Table 11. For DPO, following LLaMA-Clinic, we set $\hat{y}$ as the chosen and the student note y^- as the rejected sample. In contrast, for MEPO, we pair an edited note y^+ (chosen) with y^- (rejected), as described in Section 2.2.

3.3 Evaluation

All evaluations use a 3-LLM-jury (GPT-5.1, Claude-Sonnet-4.6, Llama-3.3-70B-Instruct) ensemble of MedHELM evaluation (Bedi et al., 2025), where each LLM-jury is given a dialogue, a generated note to evaluate, and a ground truth note generated by itself. Prior work validates MedHELM against clinician assessments, reporting a higher intraclass correlation coefficient on ACI-Bench and MEDIQA-QA than automated metrics, even exceeding the average clinician-clinician agreement. Following the MedHELM protocol, each LLM-Jury evaluates the quality of a given note and generates scores on 3 criteria (`Accuracy`, `Completeness`, `Clarity`) within the range of 1-5, using the prompt in Table 8. For the model selection, we take the average score across

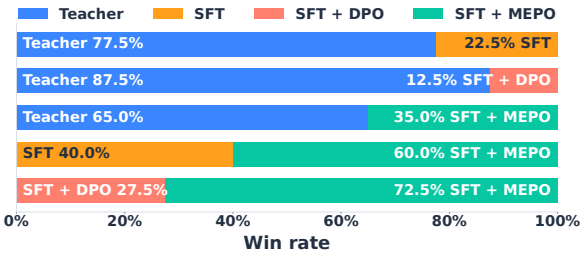

Figure 2: Pairwise win rate on 40 dialogues from test subset, judged by GPT-5-mini. Higher is better.

3 criteria and select a checkpoint with the highest score on the validation set. For the final evaluation of the test set, we report the average score and scores for each criterion.

4 Experiment Results

4.1 Main Results

Table 1 presents the MedHELM evaluation results on validation and test sets. The base student model achieves an average score of 3.87 on the test set, while the teacher model reaches 4.53, indicating a substantial performance gap. Among single-stage training methods, SFT improves the student model by +0.39, achieving an average score of 4.26. In contrast, DPO performs worse than SFT at 3.93, suggesting that directly treating teacher output as a preference may introduce a distribution mismatch. Although MEPO provides more subtle preference signals than SFT and DPO baselines, it still yields a notable improvement of +0.26 over the student under single-stage training, reaching an average score of 4.13. The multi-stage **SFT+MEPO** achieves the best overall performance, with an average score of **4.38**, outperforming both SFT and SFT+DPO. These results indicate that MEPO provides more effective supervision for clinical note generation than direct teacher imitation.

4.2 Pairwise Comparison with Baselines

We further compare the teacher, SFT, SFT+DPO, and SFT+MEPO through pairwise comparison on 40 dialogues from the ACI-Bench test subset, using the prompt in Table 10 and GPT-5-mini as a judge. As shown in Figure 2, SFT+MEPO is preferred over SFT and SFT+DPO on 60.0% and 72.5% of cases, respectively. The larger gap against SFT+DPO is consistent with our main hypothesis: MEPO improves the model without distribution shift. This trend also aligns with the dataset statistics in Table 4, where edited notes remain closer to the original student note than to the teacher note.

Source	Evidence
Dialogue	"... rhythm is irregularly irregular ... trace peripheral edema on extremities ..."
SFT	"... Atrial fibrillation is controlled on warfarin and atenolol; patient reports initial irritability from atenolol, now tolerated."
+DPO	"... INR currently therapeutic; metoprolol succinate 50 mg daily."
+MEPO	"Irregularly irregular rhythm with trace peripheral edema. Atenolol initially caused irritability but is now tolerated."

Table 2: Generated note excerpts of each student models, from **Case Study 2** in Appendix F. SFT omits dialogue-grounded exam findings, SFT+DPO introduces unsupported elaborations, while SFT+MEPO restores the missing findings without hallucination.

4.3 Comparison of Generated Notes

Manual inspection reveals a clear qualitative pattern: SFT+MEPO produces notes that are generally more complete than SFT while remaining better grounded in the source dialogue than SFT+DPO. In particular, MEPO often recovers clinically relevant details omitted by SFT without introducing unsupported elaborations. For example, in **Case Study 2** (Appendix F), the dialogue documents exam findings such as an irregularly irregular rhythm and traces peripheral edema. SFT omits these findings, while SFT+DPO introduces unsupported elaborations such as a therapeutic INR assessment and a switch from atenolol to metoprolol. In contrast, MEPO restores the dialogue-grounded exam findings without adding unsupported interpretations.

5 Conclusion

We introduce MEPO, an efficient distillation framework for clinical note generation that constructs preference pairs by minimally editing student notes. On ACI-Bench, SFT+MEPO achieves the strongest test performance, outperforming both SFT and SFT+DPO baselines. Our analysis suggests that this gain comes from a better balance between completeness and factual grounding. These results indicate that minimal-edit preference signals provide a scalable alternative to naive teacher imitation for aligning SLMs in privacy-sensitive clinical note generation settings.

Limitations

This work studies one practical setting of edit-based preference learning: minimal-edit supervision with DPO for AP note generation. Although this setup is sufficient to evaluate the core idea, it leaves several related questions open, including how the same supervision would behave with other post-training objectives such as GRPO or token-level PPO, and how well the method transfers to other note sections or teacher–student configurations. Finally, because the task involves clinical note generation, erroneous or hallucinated outputs could pose risks if used in real clinical workflows; therefore, our study focuses on methodological evaluation and assumes appropriate human oversight in any practical deployment.

Ethical Considerations

This work uses publicly available datasets that were previously released for research purposes. According to the original dataset publications, the dialogues consist of either de-identified clinical encounters or synthetic cases derived from medical documents. We did not collect any new human-subject data and therefore did not perform any additional anonymization procedures. Because the task involves generating clinical notes, erroneous or hallucinatory outputs could pose risks if used in real clinical workflows. Our work is intended for research purposes only, and any real-world deployment would require careful validation and clinician oversight.

References

Imjin Ahn, Hansle Gwon, Young-Hak Kim, Tae Joon Jun, and Sanghyun Park. 2024. Note: Notable generation of patient text summaries through efficient approach based on direct preference optimization. *arXiv preprint arXiv:2402.11882*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi, Asad Aali, and 62 others. 2025. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*.

Nathaniel Berger, Miriam Exel, Matthias Huck, and Stefan Riezler. 2024. Post-edits are preferences too. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1289–1300.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, and 1 others. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 22–30.

Gregory Finley, Erik Edwards, Amanda Robinson, Michael Brenndoerfer, Najmeh Sadoughi, James Fone, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. An automated medical scribe for documenting clinical encounters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–15.

John Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin An, Grace Zheng, and Bo Wang. 2023. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 323–334.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jiehao Liang, Haihui Yang, Shiping Gao, and Xiaojun Quan. 2025. Edit-wise preference optimization for grammatical error correction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3401–3414.

Sabine Molenaar, Lientje Maas, Verónica Burriel, Fabiano Dalpiaz, and Sjaak Brinkkemper. 2020. Medical dialogue summarization for automated reporting in healthcare. In *International conference on advanced information systems engineering*, pages 76–88. Springer.

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.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Doron Stupp, Ronnie Barequet, I-Ching Lee, Eyal Oren, Amir Feder, Ayelet Benjamini, Avinatan Hassidim, Yossi Matias, Eran Ofek, and Alvin Rajkomar. 2022. Structured understanding of assessment and plans in clinical documentation. *medRxiv*, pages 2022–04.

Hanyin Wang, Chufan Gao, Bolun Liu, Qiping Xu, Guleid Hussein, Mohamad El Labban, Kingsley Iheasirim, Hariprasad Reddy Korsapati, Chuck Outcalt, and Jimeng Sun. 2025a. Towards adapting open-source large language models for expert-level clinical note generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12084–12117, Vienna, Austria. Association for Computational Linguistics.

Yifan Wang, Runjin Chen, Bolian Li, David Cho, Yihe Deng, Ruqi Zhang, Tianlong Chen, Zhangyang Wang, Ananth Grama, and Junyuan Hong. 2025b. More is less: The pitfalls of multi-model synthetic preference data in dpo safety alignment. *arXiv preprint arXiv:2504.02193*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10:586.

A Training Details

All training was conducted with the Hugging Face Transformers library using LoRA adapters. We used LoRA with rank $r = 64$, $\alpha = 32$, and dropout 0.05, applied to the attention and MLP projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`). We used bf16 precision and a maximum sequence length of 16,384. We trained each model for 5 epochs with an effective batch size of 32. For optimization, we used a cosine learning-rate scheduler with a warmup ratio of 0.1, and selected the learning rate based on validation performance via greedy search over the range $[5 \times 10^{-8}, 1 \times 10^{-5}]$. We fixed the random seed to 42. For both data construction and final inference, we used deterministic decoding with temperature 0.0. All training and inference were performed on two RTX PRO 6000 GPUs.

B AP-Section Clinical Note Generation Implementation Details

Table 3 shows prompt template used to prompt student/teacher model to generate AP-focused clinical notes, given a dialogue.

AP Section Note Generation Prompt

```
You are a physician writing a clinical
note based on a dialogue with the patient.
Write only the "ASSESSMENT AND PLAN" part
of the note.
Only include information contained in the
dialogue.

Here's the actual dialogue between doctor
and the patient:
{{ dialogue }}
```

Table 3: Clinical note generation prompt template

C MedHELM Evaluation Details

Table 8 shows prompt template used to prompt LLM-jury to evaluate a given note.

D Minimal Edit Implementation Details

Table 9 shows prompt template used to prompt teacher model to conduct minimal-edits on student-generated notes.

E Minimal Edit Example

Dataset-level statistics of the constructed MEPO preference pairs are summarized in Table 4. Table 11 and Table 12 shows input and output sample and minimal editing process. We highlight two main observations: (1) minimal edits typically preserve proximity to the original policy note while still moving the note toward the teacher, and (2) teacher intervention scales with policy quality.

Minimal edits preserve the student distribution. A central motivation of MEPO is to avoid replacing the student note with a fully rewritten teacher note. When we compare token-level Dice overlap, the edited note remains closer to the student note than the teacher note. We find that edited notes are closer to the student notes than the teacher notes for over 98% of the time. For edit on basemodel's note, the average proximity of edited note to student note (as measured in token-level Dice) is 0.760, whereas the proximity to teacher note is only 0.649. For edits on SFT model's note,

Metric	Base	SFT
Token Dice : Student ↔ Teacher	0.486	0.665
Token Dice : Edited ↔ Teacher	0.649	0.713
Token Dice : Edited ↔ Student	0.760	0.895
Average # of Edits per Note	3.689	2.677
Average Token Change Rate	0.588	0.218
Average Token Dice Gain (to Teacher)	0.163	0.048
Macro Edit (%)	64.5	20.6
Micro Edit (%)	3.0	16.3
Positive Dice Gain Toward Teacher (%)	95.7	81.3
Edited Closer to Student Than Teacher (%)	98.3	99.3

Table 4: Dataset-level statistics of MEPO training pairs. Here, Base and SFT denotes student model from which student notes are generated. Large rewrite denotes cases with token change rate > 0.30, and near-identity denotes cases with token change rate ≤ 0.05. Teacher edits adapt to student note quality while preserving proximity to the original student note.

the same pattern holds at 0.895 versus 0.713. At the same time, the edit still moves the note toward teacher note: positive token dice gain occurs in 95.66% of base model note and 81.30% of SFT model note. This suggests that MEPO provides a distribution-preserving correction signal rather than a direct imitation target.

Teacher intervention scales with note quality. The teacher does not apply a fixed re-writing pattern. Instead, edits adapt to the quality of the student note, meaning lower-quality note consistently receives more frequent and larger edits. Base model note receives more frequent edits than SFT model note (3.689 vs. 2.677 edits/note), larger edit size as measured by token replacement rate (0.588 vs. 0.218), and larger gains as measured by increase in token dice against teacher note (0.1632 vs. 0.0477). This behavior is consistent with an adaptive correction policy.

F Representative Qualitative Examples

To support the qualitative comparison in Section 4.3, we provide three representative cases selected from the same 40 dialogue test subset used in our analysis. We manually reviewed the teacher, SFT, SFT+DPO, and SFT+MEPO outputs for all cases, and selected examples where the contrast among student variants was clear and clinically interpretable. Across cases, a consistent pattern emerges: SFT often produces shorter notes that omit clinically relevant detail, whereas SFT+DPO more often adds plausible but unsupported elaborations. SFT+MEPO typically remains between these extremes, preserving the structure of the student note while recovering grounded omissions. This case-level pattern is also consistent with the dataset-level statistics in Table 4, where edited notes remain closer to the original policy note than to the teacher note. We elaborate on our case study findings on 3 sample cases, and present grounds for our conclusion in Table 5 through Table 7, highlighting successful and erroneous generations with green and red, respectively.

Case 1: Asthma follow-up. Table 5 illustrates a typical completeness–grounding trade-off across methods. SFT captures the overall diagnosis and treatment plan, but compresses the case and omits several dialogue-grounded details that are clinically useful for longitudinal management, nocturnal rescue-inhaler use, and repeated urgent-care visits and prior ED evaluation requiring steroids. SFT+DPO moves in the opposite direction: while it includes many relevant details, it also introduces unsupported or overly specific additions, such as fixed-airflow-limitation framing and extra severity interpretation that are not clearly stated in the dialogue. In contrast, SFT+MEPO preserves the core structure of the SFT note while recovering important missing details such as trigger history, prior exacerbation burden, and rescue-inhaler usage, yielding a note that is both more complete and better grounded.

Case 2: Post–mitral valve repair follow-up. The contrast in Table 6 is mainly about factual grounding. SFT already produces a reasonable summary, but SFT+MEPO improves visit-specific fidelity by preserving more of the exam and medication context described in the dialogue, including the irregularly irregular rhythm, trace edema, and the patient's tolerance of atenolol after initial irritability. By comparison, SFT+DPO introduces several clinically plausible but unsupported details, including therapeutic INR status, a switch from atenolol to metoprolol, hypertension assessment, low-cardiac-output interpretation, and explicit adherence claims. These additions make the note sound more teacher-like, but they are not clearly grounded in the source conversation. This example highlights one of the main advantages of MEPO: it improves completeness without replacing the student note with a fully rewritten teacher-style summary.

Case 3: Acute low back pain with conversational noise. Table 7 shows a different failure mode: contamination from noisy dialogue segments. The source conversation includes partner interjections that are not part of the patient's own clinical report. SFT partially carries this noise into the note by introducing partner-related symptom descriptions, and SFT+DPO amplifies the issue further with additional partner-related and iatrogenic speculation. SFT+MEPO produces a cleaner patient-centered note that focuses on the clinically relevant findings—acute mechanical back pain after exertion, right foot paresthesia, positional pain pattern, normal dorsiflexion and pulses, and a normal X-ray—without importing distracting conversational content. This suggests that minimal-edit supervision can refine a plausible student note while resisting both omission and over-interpretation.

Summary of qualitative pattern. Across the three cases, the dominant baseline failure modes differ in a consistent way: SFT tends to under-specify, whereas SFT+DPO tends to over-specify. SFT+MEPO generally occupies a more effective middle ground by preserving the student note's local structure while selectively adding missing, dialogue-grounded clinical information. This qualitative trend is aligned with the quantitative findings in both Section 4.2 and Table 4.

Method	Generated note (verbatim example)
Dialogue (abridged)	[doctor] hey betty how are you doing [patient] i'm doing okay ... i'm always coughing and wheezing . [doctor] what normally triggers your asthma type symptoms ? [patient] when i get sick with a cold i almost always know that my asthma is going to flare . i did notice one time though that when i went to my aunt's house i had a flare afterwards and i think it's because of her cat . [doctor] how frequently does this occur ? [patient] it seems like it happens really every month or every other month for the past six months . i usually end up going to the urgent care a lot ... sometimes they'll give me steroids . one time it was so bad i actually ended up going to the emergency department . ... [doctor] how often do you feel like you need to use that ? [patient] i would say like almost everyday i feel like i'm using my inhaler ... and a couple of times a night i'll wake up ... a couple times a week i'll end up waking up needing to use my inhaler as well . [doctor] you just wake up and you take one puff or is it two puffs ? [patient] yeah i usually take two puffs . ... [doctor] can you tell me a little bit more about those er and urgent care visits ? [patient] i've had several this past year ... they almost always have to start me on prednisone . [doctor] okay like a prednisone taper pack that type of thing ? [patient] yeah ... sometimes they'll just give me like a large dose over five days ... and then sometimes they make me take less and less and less over like nine or ten days . ... [doctor] you really like to go outside and walk ... i see here that you enjoyed hiking ... [patient] yeah i'm really excited to ... i'm really looking forward to getting better so i can really start hiking again . ... [doctor] i reviewed your vitals for today ... your blood pressure is good ... your oxygen saturation is ninety nine percent on room air ... [doctor] ... reviewing the results of your pulmonary function test ... they are consistent with somebody who would be diagnosed with asthma . ...[truncated for brevity]...
SFT	**ASSESSMENT AND PLAN:** 1. Moderate intermittent asthma, supported by history of recurrent exacerbations (every 1–2 months over the past 6 months), frequent nocturnal and activity-limiting symptoms, and objective findings of diminished lung sounds with occasional expiratory wheezes. 2. Inhaled albuterol (Ventolin) is effective as needed but used daily; inadequate control with current regimen. 3. Inhaled corticosteroid (Flovent) prescribed at 110 mcg twice daily to reduce frequency of albuterol use and improve long-term control. 4. Asthma action plan to be developed to manage triggers (e.g., cat exposure, secondhand smoke at bars), guide symptom response, and optimize adherence. 5. No evidence of smoking, vaping, or other environmental exposures contributing to asthma. 6. Follow-up in 30 days or sooner if symptoms worsen.
+DPO	**ASSESSMENT AND PLAN:** 1. Moderate intermittent asthma – consistent with history of recurrent respiratory exacerbations (every 1–2 months over past 6 months), triggered by viral infections, recent exposure to household pet dander (cat), and occasional environmental smoke exposure; no evidence of fixed airflow limitation or chronic severe disease on clinical and PFT correlation. - Plan: Initiate fluticasone (Flovent) 110 mcg Inhaler twice daily to optimize long-term inflammation control, reduce daily symptoms, and decrease the frequency of emergency inhaler (Ventolin) use. - Continue PRN as-needed Ventolin use: 2 puffs every 4–6 hours for acute symptoms (cough, wheeze); instruct to use only for acute exacerbations, not routine prophylaxis. - Provide written asthma action plan tailored to patient's triggers (e.g., cat exposure, bar smoke, nocturnal symptoms) to guide step-wise management of flares and environmental avoidance strategies. 2. History of prior severe exacerbation requiring ED admission at age 8 – no recent hospitalizations; current episodes managed outpatiently with occasional high-dose prednisone courses (likely reflecting underwhelming long-term controller therapy). - Avoid routine systemic corticosteroid "taper packs" for routine exacerbations; reserve for truly uncontrolled, life-threatening episodes. - Encourage consistent daily controller use to reduce need for systemic interventions. 3. No evidence of active smoking, vaping, or marijuana use; no occupational or environmental dust/metal exposures reported.

Method	Generated note (verbatim example)
	- Negative for seasonal allergies; no hand findings suggestive of chronic hypoxia or digital changes. - Vital signs and exam reveal stable hemodynamics, normoxic (SpO2 99% RA), normotensive, afebrile, with benign cardiac and abdominal exam; mild bilateral wheezes and diminished lung sounds consistent with chronic airway hyperresponsiveness on exam. . Lifestyle goals: Patient expresses motivation to resume hiking and outdoor activity; encourage gradual re-engagement with environmental precautions (avoid smoke-filled venues, consider air filtration or masks in cat-exposed settings). - Recommend low-dose inhaled corticosteroid as foundation of long-term management; reassess in 30 days or sooner if worsening dyspnea, increased nocturnal symptoms, or >2 rescue inhaler uses/week.
+MEPO	**ASSESSMENT AND PLAN:** Patient has moderate intermittent asthma, supported by history of recurrent respiratory symptoms (cough, wheezing) occurring monthly or every other month over the past six months, triggered by viral infections and exposure to a cat at her aunt's house. She reports frequent use of albuterol (Ventolin) daily for symptom relief, with two puffs at night several times a week. She has one prior hospitalization for asthma exacerbation at age 8 and multiple recent ER/urgent care visits requiring corticosteroids (prednisone or taper packs). No history of smoking, vaping, or marijuana use. Pulmonary function tests confirm asthma diagnosis. Physical exam reveals diminished lung sounds with occasional expiratory wheezes bilaterally, no clubbing, and normal oxygenation. **Plan:** 1. Initiate fluticasone (Flovent) 110 mcg twice daily as prescribed inhaler for long-term control. 2. Continue as-needed albuterol (Ventolin) 2 puffs every 4–6 hours p.r.n. for acute exacerbations; instruct to use only when needed and reassess after each dose. 3. Develop and provide asthma action plan to guide management of triggers (e.g., cat exposure, bar smoke) and symptom escalation. 4. Follow up in 30 days or sooner if symptoms worsen. 5. Encourage avoidance of known triggers and consider environmental modifications as part of action plan.

Table 5: Case study 1: verbatim generated notes from SFT, DPO, and MEPO on the same dialogue(D2N119-aci)

Method	Generated note (verbatim example)
Dialogue	[doctor] next patient is melissa sanchez ... status post mitral valve repair done on 8/3/2020 . ecg taken on 12/26/2020 reveals atrial fibrillation with a controlled ventricular response . t-wave inversion anteriorly . compared to the previous study , there are no significant changes ... ... [patient] pretty good . i'm definitely feeling better ... i was having a rough time before surgery , but i think i'm overall pretty good now though . [doctor] are you having any new symptoms right now ? [patient] no , not really . but , you know , i'm still getting some chest pains sometimes , and my breathing gets shallow ... ... [doctor] are you taking your medications regularly ? [patient] i've been trying to keep up with that ... i have a reminder app , so i do pretty well . [doctor] okay . that's great . and so , you're taking coumadin , right ? [patient] yeah , also lasix and the atenolol . [doctor] are you having any side effects from the medications at all ? [patient] not really . uh , i notice that the atenolol is making me irritable in the beginning , but i guess i've gotten used to it , does n't seem to bother me as much now . ... [doctor] head and neck no jvd detected . [doctor] lungs have reduced breath , but auscultation and percussion are clear . [doctor] rhythm is irregularly irregular . [doctor] s1 slightly accentuated , no s3 . [doctor] trace peripheral edema on extremities . ... [doctor] you're taking coumadin four milligrams , lasix four milligrams a day , and the atenolol ... [patient] yes , the 50 milligrams every day . ...[truncated for brevity]...
SFT	**ASSESSMENT AND PLAN:** Melissa Sanchez is a post-operative status following mitral valve repair (8/3/2020) with ongoing atrial fibrillation and mitral regurgitation. She is doing well overall, with no acute cardiac symptoms at this time. She reports intermittent chest discomfort (episodes lasting minutes, occurring 1-2 times/month), exertional dyspnea, and occasional shallow breathingsymptoms she manages by reducing exertion and pacing herself. No signs of heart failure (no JVD, clear lungs, trace peripheral edema). Atrial fibrillation is controlled on anticoagulation with warfarin 4 mg daily and on beta-blocker atenolol 50 mg daily; no side effects reported beyond initial irritability from atenolol, now tolerated. ECG shows persistent atrial fibrillation with no new changes since prior study; T-wave inversion anteriorly is stable. Diabetes, asthma, and recurrent chest discomfort with negative CAD workup are noted in history. Plan: Continue current medications (warfarin, lasix, atenolol) without adjustment. Follow up in 6-9 months for cardiac status. Advise continued activity tolerance and symptom monitoring.
+DPO	**ASSESSMENT AND PLAN:** 1. Status post mitral valve repair (8/3/2020) - doing well with stable cardiac recovery; no signs of valve dysfunction or congestive heart failure on exam. 2. Permanent atrial fibrillation, currently in controlled, irregularly irregular rhythm on rhythm strip and clinically; INR currently therapeutic (on Coumadin 4 mg daily), with no acute coagulopathy noted. 3. Mild t-wave inversion anteriorly - stable since prior ECG (12/26/2020); no new ischemic or structural changes identified. 4. Chronic coronary artery disease ruled out by prior negative cardiac workup; no new exertional dyspnea or anginal symptoms reported beyond occasional, intermittent, non-cardiac-explanatory chest discomfort. 5. Chronic obstructive pulmonary disease / asthma - contributing to subjective dyspnea and shallow breathing; improved with activity modification and no acute respiratory infection noted. 6. Type 2 diabetes - stable, no recent HbA1c or glucose logs reviewed today; continue routine monitoring. 7. Hypertension - adequately controlled on metoprolol succinate 50 mg daily; no orthostasis or hypotension on exam. 8. Mild trace peripheral edema noted on exam - likely related to chronic venous insufficiency and low cardiac output state post-MVR; no pitting edema, no dyspnea at rest or with exertion, well-controlled with Lasix 4 mg daily. 9. Patient demonstrates good medication adherence via pill reminder app. **Plan:** - Continue current medications: warfarin 4 mg daily (INR-stable), metoprolol succinate 50 mg daily, furosemide 4 mg daily. - Maintain rhythm control for atrial fibrillation with anticoagulation; no indication for rate control or electrical cardioversion at this time.

Method	Generated note (verbatim example)
	- Advise patient to remain physically active within symptom limits; no formal activity clearance needed. - No urgent cardiology or outpatient imaging indicated. - Follow-up in 6-9 months for repeat INR, vital signs, and clinical reassessment. - Patient may consider outpatient echocardiogram in 12-24 months for routine surveillance, pending clinical status. - Encourage self-monitoring of chest discomfort patterns; no immediate cardiac alarm features present. - Continue routine diabetes and asthma management with primary care.
+MEPO	**ASSESSMENT AND PLAN:** Melissa Sanchez is recovering well following mitral valve repair performed on 8/3/2020. She continues to have intermittent chest discomfort and occasional shallow breathing, but symptoms are mild, self-limited, and improved with activity modification. Atrial fibrillation with controlled ventricular response is present, consistent with prior ECG (12/26/2020); no new changes noted. T-wave inversion anteriorly is stable. She has a history of mitral regurgitation, atrial fibrillation, diabetes, asthma, and recurrent chest discomfort without evidence of coronary artery disease. Medications include warfarin 4 mg, furosemide 4 mg, and atenolol 50 mg daily, with no significant side effects except transient irritability from atenolol, now tolerated. Physical exam reveals irregularly irregular rhythm on auscultation, S1 slightly accentuated, no S3, no JVD, clear lungs, and trace peripheral edema. No abdominal or lower extremity tenderness. Plan: Continue current medications. Follow up in 6-9 months for cardiac status. Advise continued activity modification and symptom monitoring.

Table 6: Case study 2: verbatim generated notes from SFT, DPO, and MEPO on the same dialogue (D2N103-virtscribe).

Method	Generated note (verbatim example)
Dialogue (abridged)	[doctor] so bryan it's nice to see you again in the office today what's going on [patient] i was in my yard yesterday and i was raking leaves and i felt fine and then when i got into my house about two hours later my back started tightening up and i started getting pins and needles in my right foot [doctor] alright have you ever had this type of back pain before [patient] i had it once about three years ago but it went away after a day [doctor] okay and did you try anything for the pain yet did you take anything or have you have you tried icing [patient] put some ice on it and i tried two advils and it did n't help [doctor] okay does it get better when you're laying down sitting standing [patient] yeah i think the only thing that helped was a hot shower and it feels a little better when i sit [doctor] okay alright so you said you were raking leaves and now it's been kind of that weather outside for cleaning out the yard have you been doing a lot of yard work lately [patient] i i i normally do n't do any physical activity [doctor] yeah i hear you i make a i make my husband do all the yard work too so alright let me take a look at that low back if if you bend over does that hurt [patient] it hurts a little bit when i bend over yes [doctor] okay how about when you stand back up [patient] no pain [doctor] okay so no pain on extension some pain on flexion how about when i push right here on your back [patient] yeah that hurts [doctor] okay so some pain to palpation right around the l5 can i have that [patient] what happens when family members chime in like i'm doing right now because i actually think it was it hurts more when he stands up because i've watched him bend over but it's when he stands up that it really hurts ...[truncated for brevity]...
SFT	**ASSESSMENT AND PLAN:** Patient presents with acute low back pain following physical exertion (raking leaves) and associated right foot paresthesia, consistent with mechanical low back pain. History reveals prior similar episode three years ago, resolved spontaneously. Physical exam shows localized tenderness at L5, pain with flexion but not extension, and mild improvement with heat and sitting. Dorsiflexion is normal, pulses are equal, and no red flags on neurologic exam. X-ray is normal. No evidence of red flags for serious pathology. Pain is not improved by NSAIDs or ice. Patient reports new or worsened symptoms despite minimal prior activity, with partner also reporting similar symptoms. No signs of acute nerve root compression or red flags for urgent imaging. **Plan:** 1. Prescribe meloxicam for anti-inflammatory and analgesic effect. 2. Refer to physical therapy for core strengthening, flexibility, and targeted back exercises. 3. Advise activity modification and avoidance of heavy lifting or prolonged standing. 4. Instruct patient to follow up in two weeks if no improvement, at which time an MRI may be considered. 5. Reassure that MRI is not indicated at this time given lack of red flags and normal X-ray; avoid premature imaging unless symptoms persist.
+DPO	**ASSESSMENT AND PLAN:** 1. Acute low back strain/sprain likely mechanical in origin, triggered by recent sedentary lifestyle followed by abrupt physical exertion (raking leaves); associated right radicular-type symptomatology with isolated right foot paresthesia, consistent with transient, non-red flag neurological involvement. 2. No red flags for cauda equina, spinal cord, or significant structural pathology per prior lumbar X-ray; no neurovascular deficits beyond localized paresthesia and preserved motor strength/pulses. 3. Pain localized to L5 region, exacerbated by flexion and relieved by extension, positional relief with hot showers, and milder activity; no improvement with NSAIDs or simple ice. 4. Significant new onset of persistent back pain (~6 hours) with associated functional limitation; patient and partner report worsening symptoms over prior weeks with minimal prior activity, raising concern for prolonged, unrecognized mechanical stressor. **Plan:** - Strict activity modification: avoid heavy lifting, prolonged standing/bending, and high-impact activities; encourage gentle ambulation. - Initiate meloxicam 7.5 mg PO daily for 7 days for anti-inflammatory/pain modulation; avoid NSAIDs if recent surgery in partner—clarify timing and indication for urgent imaging. - Referral to physical therapy for core stabilization, lumbar bracing education, and targeted lumbar flexibility/strengthening protocol; focus on posture, body mechanics, and gradual return to activity. - Advise patient to return or message clinic if: a) Pain worsens or fails to improve within 2 weeks b) New motor weakness, bowel/bladder dysfunction, or worsening radicular symptoms c) If partner's symptoms persist despite conservative measures—consider concurrent evaluation for possible iatrogenic or chronic mechanical etiology - MRI of lumbar spine deferred unless clinically deteriorating or non-responsive to conservative management; avoid imaging solely for chronic, non-red-flag musculoskeletal pain unless indicated by clinical progression.

Method	Generated note (verbatim example)
	- Reassess in 2 weeks or sooner if symptoms progress; encourage open communication regarding partner's concerns regarding prior surgery and current alignment of care.
+MEPO	**ASSESSMENT AND PLAN:** Patient presents with acute low back pain following physical exertion (raking leaves) and subsequent development of right foot paresthesias, consistent with mechanical low back pain. History reveals prior similar episode three years ago, transient and resolved. Pain is localized to L5 region, exacerbated by flexion and relieved by hot showers and sitting; no improvement with NSAIDs or ice. Dorsiflexion is normal, pulses equal, and no red flags on exam. X-ray of lumbar spine is normal. Clinical picture is consistent with acute lumbar sprain/strain. No evidence of red flags for cauda equina or significant neurologic deficit. **Plan:** 1. Prescribe meloxicam for anti-inflammatory and analgesic effect. 2. Refer to physical therapy for core strengthening, lumbar stabilization, and targeted stretching exercises. 3. Advise activity modification and avoidance of heavy lifting or prolonged standing/bending. 4. Instruct patient to follow up in two weeks if no improvement; if persistent or worsening, consider MRI to evaluate for disc pathology or other structural cause. 5. Reassured patient that MRI is reserved for non-responsive cases; no immediate imaging indicated at this time.

Table 7: Case study 3: verbatim generated notes from SFT, DPO, and MEPO on the same dialogue (D2N107-aci).

MedHELM Evaluation Prompt

```
You are a medical expert tasked with evaluating the quality of a generated clinical note based
on a doctor-patient conversation.
Your goal is to assess how well the note captures the clinical information from the
conversation and compare it to the reference note (gold standard) in terms of accuracy,
completeness and clarity.

The conversation will be provided in these tags:
<conversation>
</conversation>

The generated note will be provided in these tags:
<response>
</response>

The reference note will be provided in these tags:
<gold_response>
</gold_response>

Carefully review the <response> based on the <conversation> and compare it to the
<gold_response> when needed.

For each of the following criteria, rate the response on a scale of 1 to 5 (1 = very poor, 5 =
excellent),
and provide a short justification for your score.

Evaluation Criteria:
Accuracy (1-5)
- Does the note provide correct clinical information based on the conversation?

Completeness (1-5)
- Does the note include all important medical details from the conversation?

Clarity (1-5)
- Is the note written clearly and organized in a standard clinical format for clinicians

Output Format:
Output the evaluation as a single valid JSON object matching the following structure:
{
    "accuracy": {
        "score": 0,
        "explanation": "Explain why this score was given."
    },
    "completeness": {
        "score": 0,
        "explanation": "Explain why this score was given."
    },
    "clarity": {
        "score": 0,
        "explanation": "Explain why this score was given."
    }
}

Ensure the output is valid JSON:
- Use **double quotes** ("") for all keys and string values.
- When quoting text or sections inside the explanations, use escaped double quotes (\") to
  maintain valid JSON formatting.
- Do not include any additional information in the output.

<conversation>\n{{ context }}\n</conversation>
<response>\n{{ response }}\n</response>
<gold_response>\n{{ gold_response }}\n</gold_response>
```

Table 8: MedHELM LLM-jury evaluation prompt template

Minimal Edit Prompt

```
Given (1) the dialogue context between doctor and patient, (2) a reference clinical note that
is relatively accurate but not necessarily perfect  it may omit certain details or contain
minor imprecisions, and (3) a candidate clinical note that may contain more errors or
inaccuracies compared to the reference, your task is to minimally correct and improve the
candidate clinical note by grounding your revisions in the actual dialogue content. The
reference clinical note should be used only as a supplementary guide, not as the ground truth.
The goal is to produce a corrected clinical note that faithfully reflects what was discussed in
the dialogue, as a real-world medical consultant would write it.

## Input

You will receive:
    1. **dialogue**: The dialogue context between doctor and patient
    2. **reference clinical note**: A pre-validated clinical note that is relatively accurate
    and faithful to the dialogue, but not necessarily perfect  it may contain minor
    imprecisions or omissions.
    3. **candidate clinical note**: A clinical note that may contain factual inaccuracies,
    medical hallucinations, or critical omissions, and is generally less accurate than the
    reference clinical note.

## Output

Follow these detailed steps and output your reasoning for each step wrapped for each respective
XML tag below:
    1. You should think and provide your reasoning under <reasoning>[...]</reasoning> on how
    the candidate clinical note falls short compared to the dialogue content  such as factual
    inaccuracies, hallucinations, or critical omissions relative to what was actually
    discussed. The reference clinical note may be consulted as a supplementary aid to identify
    potential gaps, but your reasoning should be primarily **grounded in the dialogue itself**.
    2. Next, for all indentified issue, you must include the original text with
    <before>[...]</before>, the corrected version within <after>[...]</after>, and the specific
    reasoning or dialogue-based evidence for the change within <evidence>[...]<evidence>. You
    must preserve the original structure of the candidate clinical note and ensure the edit
    distance is as low as possible while maintaining strict grounding in the dialogue content.
    3. Finally, output the complete, integrated version of the clinical note after all
    modifications have been applied in <note>[...]</note>. This final note should be cohesive
    medical document that faithfully reflects what was discussed in the dialogue, as
    real-world medical consultant would write it.

Follow strictly the format below:
<reasoning> **Based on the dialogue**, explain the factual inaccuracies, hallucinations, or
critical omissions in the candidate clinical note. The reference clinical note may be used only
as a supplementary guide, not as ground truth. </reasoning>
<modified> Minimally modify the candidate clinical note to address the issues identified in
your reasoning and ensure every modification is directly supported by the dialogue content.
<evidence> the clinical reasoning for the change, citing the specific part of the dialogue that
supports this correction </evidence>
    <before> the exact original segment from the candidate clinical note that contains an error
    or omission </before>^^I
    <after> the corrected version of the segment, ensuring it is minimally modified yet
    factually accurate based on the dialogue </after>
</modified>
<note> the definitive clinical note that resolves all discrepancies, ensuring the final
documentation is clear, comprehensive, and factually precise based on the dialogue.
</note>

<dialogue>\n{{ dialogue }}\n</dialogue>
<reference>\n{{ reference }}\n</reference>
<candidate>\n{{ candidate }}\n</candidate>
```

Table 9: Minimal Edit prompt template

Pairwise Comparison Prompt

```
You are a medical expert tasked with evaluating the quality of a pair of clinical notes based
on a doctor-patient dialogue. Given (1) the dialogue context between doctor and patient, (2)
clinical note A, and (3) clinical note B, Your task is to assess how well each note captures
the clinical information from the dialogue, compare them in terms of accruacy, completeness
and clarity, and determine which of the two notes serves as a better clinical note for doctor.

## Input

You will receive:
  1. *dialogue*: The dialogue context between doctor and patient.
  2. *note A*: A clinical note covering the Assessment and Plan derived from the given dialogue.
  3. *note B*: A clinical note covering the Assessment and Plan derived from the given dialogue.

## Criteria

Use the following criteria to judge the quality of each clinical note:
  - *Accuracy*: Does the note provide correct clinical information grounded on the dialogue?
  - *Completeness*: Does the note include all important medical details from the conversation?
  - *Clarity*: Is the note written clearly and organized in a standard clinical format for
  clinicians?

## Output

Follow these detailed steps and output your reasoning for each step wrapped for each respective
XML tag below:
  1. Your should think and provide your reasoning under <reasoning>[...]</reasoning> on how
  each note aligns with the specific requirements of following the criteria above.
  2. Next, compare each notes directly, identifying the key differences under
  <compare>[...]</compare>.
  3. Based on the comparison, provide your final verdict under <verdict>[...]</verdict>. If one
  is better than the other, output either A or B. If both note are comparable in quality,
  output T.

Follow strictly the format below:
<reasoning> Evaluate the reasoning individually for the two clinical notes based on the
criteria
  note (A): output the reasoning here
  note (B): output the reasoning here
</reasoning>
<compare> Compare note A and note B based on the critera </compare>
<verdict> Based on your reasoning and comparison, output your final verdict in one of the
following options: A, B, or T </verdict>

<dialogue>\n{{ dialogue }}\n</dialogue>
<note_a>\n{{ note_a }}\n</note_a>
<note_b>\n{{ note_b }}\n</note_b>
```

Table 10: Pairwise comparison prompt template

Stage	Content
dialogue Evidence	[Doctor]: Good morning, Mr. [Patient], how are you feeling today? [Patient]: I'm not doing so well, doctor. I've been having trouble breathing for the past three days. I also have a cough and a low-grade fever. [Doctor]: I see. Can you tell me more about your breathing problems? [Patient]: It feels like I can't get enough air. I'm also coughing a lot, and it's getting worse when I cough. [Doctor]: Do you have any chest pain? [Patient]: Yes, I have some chest pain on both sides. It doesn't radiate to my arms or jaw, but it gets worse when I cough. [Doctor]: How long have you been having these symptoms? [Patient]: I first had them about 10 months ago, but they've been getting worse since then. [Doctor]: I see. Do you have any history of medical problems? [Patient]: Yes, I have diabetes, high blood pressure, COPD, and coronary artery disease. [Doctor]: Okay. And have you had any surgeries in the past? [Patient]: Yes, I've had a CABG, an appendectomy, and a tonsillectomy. [Doctor]: Do you have any allergies? [Patient]: No, I don't have any drug allergies. [Doctor]: What medications are you taking? [Patient]: I take Advair twice a day, Lisinopril once a day, Lopressor twice a day, Aspirin once a day, Plavix once a day, Multivitamins, Feso4 once a day, and Colace once a day. [Doctor]: Okay. I'm going to examine you now. [Doctor]: (Examines the patient) Your lungs sound clear, but I can hear some wheezing. Your heart sounds normal. Your blood pressure is a little high, but your pulse is regular. [Doctor]: Based on your symptoms and my examination, I think you have an exacerbation of your COPD. I'm going to prescribe you some antibiotics and steroids to help clear up the infection. I also want you to start using a nebulizer with albuterol three times a day. [Patient]: Okay, I'll do that. [Doctor]: I want you to follow up with me in a week to see how you're doing. [Patient]: I will. Thank you, doctor.
Teacher raw output	<reasoning> The candidate clinical note is largely accurate and closely aligned with the dialogue, but it contains one critical inaccuracy: it states that the patient has a 'productive cough.' In the dialogue, the patient describes having a cough but never mentions producing sputum or any characteristics of the cough being "productive." The patient says: "I've been having trouble breathing... I also have a cough and a low-grade fever," and later, "I'm also coughing a lot." There is no mention of phlegm, mucus, or color of secretions. Therefore, labeling the cough as "productive" is an unwarranted inference and constitutes a medical hallucination. Additionally, while the candidate note correctly lists all medications, it does so in a way that slightly deviates from the exact phrasing used by the patient (e.g., "Feso4" vs. "ferrous sulfate"), but this is a minor terminology variation acceptable in clinical documentation. The reference note also does not mention 'productive cough,' which supports that this detail was not stated. The rest of the candidate note including the assessment of COPD exacerbation, wheezing on exam, comorbidities, medication list, and plan is factually grounded in the dialogue. The only substantive error is the unsupported descriptor 'productive cough.' Thus, the minimal necessary correction is to remove the word "productive" to avoid introducing a hallucinated detail not present in the dialogue. </reasoning> <modified> <evidence>The patient states: "I've been having trouble breathing for the past three days. I also have a cough and a low-grade fever." and "I'm also coughing a lot." There is no mention of sputum, phlegm, or any characteristic indicating the cough is productive. Therefore, describing it as "productive" is an unsupported inference and must be removed.</evidence> <before>Patient is experiencing an exacerbation of COPD, likely triggered by superimposed infection, given persistent dyspnea, productive cough, low-grade fever, chest pain worsened by coughing, and history of COPD.</before> <after>Patient is experiencing an exacerbation of COPD, likely triggered by superimposed infection, given persistent dyspnea, cough, low-grade fever, chest pain worsened by coughing, and history of COPD.</after> </modified>

Table 11: Minimal Edit Example 1 - dialogue and Teacher Output

Stage	Content
Student Note (*Rejected*)	ASSESSMENT AND PLAN: Patient is experiencing an exacerbation of COPD, likely triggered by superimposed infection, given persistent dyspnea, productive cough, low-grade fever, chest pain worsened by coughing, and history of COPD. Physical exam reveals wheezing; lungs otherwise clear, heart sounds normal. Chronic comorbidities include diabetes, hypertension, coronary artery disease, and prior CABG. Plan: - Initiate antibiotics for presumed respiratory infection. - Prescribe systemic corticosteroids to reduce airway inflammation. - Start nebulizer treatments with albuterol three times daily for bronchodilation. - Continue current medications: Advair, Lisinopril, Lopressor, aspirin, Plavix, multivitamin, ferrous sulfate, and colace. - Follow up in one week to assess response to therapy.
Edited Note (*Chosen*)	ASSESSMENT AND PLAN: Patient is experiencing an exacerbation of COPD, likely triggered by superimposed infection, given persistent dyspnea, cough, low-grade fever, chest pain worsened by coughing, and history of COPD. Physical exam reveals wheezing; lungs otherwise clear, heart sounds normal. Chronic comorbidities include diabetes, hypertension, coronary artery disease, and prior CABG. Plan: - Initiate antibiotics for presumed respiratory infection. - Prescribe systemic corticosteroids to reduce airway inflammation. - Start nebulizer treatments with albuterol three times daily for bronchodilation. - Continue current medications: Advair, Lisinopril, Lopressor, aspirin, Plavix, multivitamin, ferrous sulfate, and colace. - Follow up in one week to assess response to therapy.

Table 12: Minimal Edit Example 1 - Note Comparison

Stage	Content
dialogue Evidence	[Doctor]: Good morning, Mrs. [Patient], I'm Dr. [Doctor]. I understand you're here today for a consultation regarding a left breast mass. [Patient]: Yes, that's right. I've been having some issues with my left breast lately. [Doctor]: I see. Can you tell me a little bit more about what you've been experiencing? [Patient]: Well, I first noticed a small lump in my left breast a few weeks ago. It's about the size of a pea, and it's located at the 7 o'clock position. It's not painful, but it's definitely noticeable. [Doctor]: I see. Have you noticed any other changes in your breast, such as nipple discharge or skin dimpling? [Patient]: No, I haven't noticed anything like that. [Doctor]: That's good to hear. Have you had any recent breast imaging, such as a mammogram or ultrasound? [Patient]: Yes, I had a mammogram and an ultrasound a few weeks ago. The results showed a 1.5 cm lesion in my left breast. [Doctor]: I see. And did you have a biopsy of the lesion? [Patient]: Yes, I did. The biopsy results came back positive for infiltrating ductal carcinoma. [Doctor]: I'm sorry to hear that. Can you tell me a little bit about your medical history? [Patient]: Well, I've had a few surgeries in the past, including a benign breast biopsy in 1972, a laparotomy in 1981, 1982, and 1984, a right oophorectomy in 1984, and a ganglion cyst removal of the hand in 1987. [Doctor]: I see. And what about your family history? [Patient]: My mother had breast cancer at age 48. She also had mesothelioma from possible asbestosis or asbestos exposure. [Doctor]: That's important information. Now, I'd like to examine your breast. [The doctor performs a breast exam.] [Doctor]: The exam reveals a 1.5 cm relatively mobile focal palpable mass in the inferomedial left breast at the 7 o'clock position. It's not fixed to the underlying pectoralis muscle. There are no nipple retractions or skin dimpling. There is some ecchymosis from the recent biopsy. There is no axillary adenopathy. The remainder of the left breast is without abnormality. The right breast is also without abnormality. The axillary areas are negative for adenopathy bilaterally. [Patient]: What does this all mean? [Doctor]: Well, based on your symptoms, the biopsy results, and the physical exam, it appears that you have invasive ductal carcinoma of the left breast. [Patient]: Oh no. What does that mean? [Doctor]: Invasive ductal carcinoma is the most common type of breast cancer. It starts in the milk ducts of the breast and can spread to other parts of the body if it's not treated. [Patient]: Is it curable? [Doctor]: In many cases, yes. The type of treatment you receive will depend on the stage of the cancer and your overall health. [Patient]: What are my treatment options? [Doctor]: Well, we can discuss that in more detail once we have a better understanding of the stage of your cancer. But generally speaking, the main treatment options for breast cancer include surgery, chemotherapy, radiation therapy, and targeted therapy. [Patient]: I see. [Doctor]: I know this is a lot to take in. I want to assure you that we're here to help you through this. We'll work with you to develop a treatment plan that's right for you. [Patient]: Thank you, Doctor. [Doctor]: You're welcome. Now, I'd like to refer you to a medical oncologist and a radiation oncologist. They'll be able to provide you with more information about your treatment options. [Patient]: Okay. [Doctor]: I'll also give you some brochures about breast cancer and its treatment. Please take some time to read them over. [Patient]: I will. [Doctor]: And of course, if you have any questions, please don't hesitate to ask. [Patient]: Thank you, Doctor. I appreciate your help. [Doctor]: You're welcome.

Table 13: Minimal Edit Example 2 - dialogue

Stage	Content
Teacher raw output	<reasoning> The candidate clinical note is largely accurate but contains two critical issues that deviate from the dialogue content: 1. **Incorrect implication of "possible asbestosis or asbestos exposure" as a confirmed diagnosis in the family history**: The patient states that her mother had mesothelioma from possible asbestosis or asbestos exposure, which is speculative. The candidate note incorrectly presents this as a factual part of the family history (mesothelioma from possible asbestosis or asbestos exposure), which could mislead future providers into treating it as a confirmed etiology or hereditary risk factor. The dialogue clearly frames it as the patients report of a possible cause, not a confirmed diagnosis. The reference note correctly omits this speculative detail. 2. **Unwarranted addition of genetic counseling if desired by the patient**: The dialogue contains no mention of genetic testing, BRCA status, or patient interest in genetic counseling. While genetic counseling may be clinically appropriate in a patient with early-onset familial breast cancer, the candidate note introduces this as a planned action without any basis in the dialogue. The reference note also does not include this, and since our task is to ground revisions strictly in the dialogue, this is a hallucination. All other elements in the candidate note the mass location, size, biopsy result, absence of skin/nipple/axillary findings, past surgeries, and mothers breast cancer are accurate and consistent with the dialogue. The plan items regarding referrals and brochures are also correct and align with the dialogue. Therefore, only two minimal corrections are needed: (1) remove the speculative etiology of mesothelioma, and (2) remove the unsupported recommendation for genetic counseling. </reasoning> <modified> <evidence>The patient states: 'My mother had breast cancer at age 48. She also had mesothelioma from possible asbestosis or asbestos exposure.' The phrase 'from possible asbestosis or asbestos exposure' is the patients speculation, not a confirmed medical diagnosis. Including it as part of the family history in the note implies a causal or diagnostic certainty not present in the dialogue. The reference note correctly omits this speculative detail, and the candidate note should too.</evidence> <before>Family history includes breast cancer in her mother at age 48 and mesothelioma from possible asbestosis or asbestos exposure.</before> <after>Family history includes breast cancer in her mother at age 48.</after> </modified> <modified> <evidence>Nowhere in the dialogue does the patient express interest in genetic testing, nor does the doctor mention genetic risk, BRCA status, or recommend genetic counseling. While this may be clinically reasonable, the candidate note introduces it as a planned action without dialogue support. The reference note also omits this, and our task requires strict grounding in the dialogue. Therefore, this item must be removed.</evidence> <before>4. Arrange for genetic counseling if desired by the patient.</before> <after></after> </modified>

Table 14: Minimal Edit Example 2 - Teacher Output

Stage	Content
Student Note (*Rejected*)	ASSESSMENT AND PLAN: The patient presents with a left breast mass, confirmed by biopsy to be infiltrating ductal carcinoma. The mass is 1.5 cm in size, located in the inferomedial left breast at the 7 o'clock position, and is not fixed to the underlying pectoralis muscle. There are no nipple retractions, skin dimpling, or axillary adenopathy. The patient has a history of benign breast biopsy, laparotomy, right oophorectomy, and ganglion cyst removal. Family history includes breast cancer in her mother at age 48 and mesothelioma from possible asbestosis or asbestos exposure. PLAN: 1. Refer to a medical oncologist and a radiation oncologist for further evaluation and development of a treatment plan. 2. Provide the patient with brochures about breast cancer and its treatment. 3. Schedule a follow-up appointment to discuss the results of any additional imaging or staging studies. 4. Arrange for genetic counseling if desired by the patient.
Edited Note (*Chosen*)	ASSESSMENT AND PLAN: The patient presents with a left breast mass, confirmed by biopsy to be infiltrating ductal carcinoma. The mass is 1.5 cm in size, located in the inferomedial left breast at the 7 o'clock position, and is not fixed to the underlying pectoralis muscle. There are no nipple retractions, skin dimpling, or axillary adenopathy. The patient has a history of benign breast biopsy, laparotomy, right oophorectomy, and ganglion cyst removal. Family history includes breast cancer in her mother at age 48 PLAN: 1. Refer to a medical oncologist and a radiation oncologist for further evaluation and development of a treatment plan. 2. Provide the patient with brochures about breast cancer and its treatment. 3. Schedule a follow-up appointment to discuss the results of any additional imaging or staging studies.

Table 15: Minimal Edit Example 2 - Note Comparison