MedEvo: Medical Reasoning Evolution through Tag-Guided Navigation

Anonymous authors
Paper under double-blind review

Abstract

Medical multimodal reasoning has advanced through teacher-generated rationale distillation, verifier- or reward-guided learning, and on-policy supervision over student-generated rollouts. Although these paradigms can exploit mistakes through rejection, reward, preference, process, or token-level signals, explicitly recovering valid reasoning states from failed medical trajectories and converting localized corrections into reusable, student-relative supervision remains underexplored. Moreover, general-domain correction mechanisms do not directly address medical failures grounded in modality-specific visual evidence, anatomical context, evidence integration, and differential diagnosis. We propose MedEvo, a failure-conditioned framework for evolving medical multimodal reasoning through localized trajectory repair and feedback-driven data evolution. At the trajectory level, MedEvo uses step-level uncertainty to propose a candidate failure region, employs a gold-blind auditor to identify a safe rollback boundary and generate a taxonomy-guided minimal recovery hint, and lets the original student preserve the valid prefix while regenerating only the erroneous suffix. Repair-aware learnability filtering then retains answer-correct, locally learnable trajectories that remain compatible with the student policy. At the data level, a fine-grained medical taxonomy characterizes checkpoint-specific capability gaps and guides the selection of targeted supervision for the next evolution round. Using supervised fine-tuning alone, MedEvo improves Qwen3.5-9B on MedXpertQA-MM from 47.40% to 54.50% and raises its seven-benchmark mean from 65.57% to 69.12%, a gain of 3.55 percentage points. Experiments on 9B, 35B, and 397B models show gains in mean benchmark accuracy across scales, while round-wise comparisons examine feedback-driven data evolution. These results establish localized failure repair and structured feedback as a practical step toward recursive self-improvement in medical reasoning.

1 Introduction

Medical multimodal reasoning requires a model to connect visual findings, anatomical structure, disease knowledge, and a clinical decision. Medical vision-language models have broadened access to these capabilities, yet expert-level questions remain challenging (Li et al., 2023; Moor et al., 2023; Zuo et al., 2025). An incorrect observation or an unsupported inference can redirect the rest of a reasoning chain, even when its earlier steps are useful. Learning to recover at such points is therefore central to improving medical reasoning.

Most reasoning supervision emphasizes complete successful trajectories: a student learns from a teacher solution or from its own answer-correct rollouts. Failed trajectories are often discarded or replaced as a whole. This loses two forms of information. Within a trajectory, an evidence-grounded prefix reveals what the student already understands. Across trajectories, the distribution of failures over medical tasks reveals which capabilities need further training. Treating all failures alike obscures both the local reasoning operation that needs correction and the medical knowledge or perceptual skill that needs additional supervision.

These observations motivate two research questions. *How can we locate and navigate a local error without discarding the valid reasoning that precedes it? How can failures across many questions be converted into medical capability feedback that guides the next round of data selection?* The first

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

question concerns how to construct a useful learning example; the second concerns which examples the model should learn from next.

We introduce MedEvo, a framework for medical reasoning evolution through tag-guided navigation. Its inner process localizes a failed reasoning step, uses observable medical context to form a minimal recovery hint, and lets the same student continue from an evidence-supported prefix. Quality gates retain useful, learnable trajectories. Its outer process evaluates the model, aggregates errors by medical tags, and directs new data construction toward capability gaps. Figure 1 illustrates the local navigation interface; Figure 2 connects this interface to the feedback loop. Together, the two processes turn individual failures into supervision and population-level failures into a learning curriculum.

Our contributions are threefold. (1) We develop tag-guided local navigation that preserves a student's valid prefix and supplies a targeted recovery operation instead of a complete teacher answer. (2) We formulate an *evaluate–aggregate–select–train* loop that links medical capability feedback to the next training dataset. (3) We evaluate first-round training on multiple medical multimodal benchmarks and analyze its components; the round-wise comparisons examine the additional value of tag feedback, separately from extra data and training initialization.

MedEvo minimizes teacher intervention: the teacher provides only a brief, error-conditioned cue identifying what to reconsider, while the student performs the corrective reasoning itself. The student-generated correct and repaired trajectories then form supervision for the next model, yielding teacher-guided self-distillation that reuses the student's own reasoning.

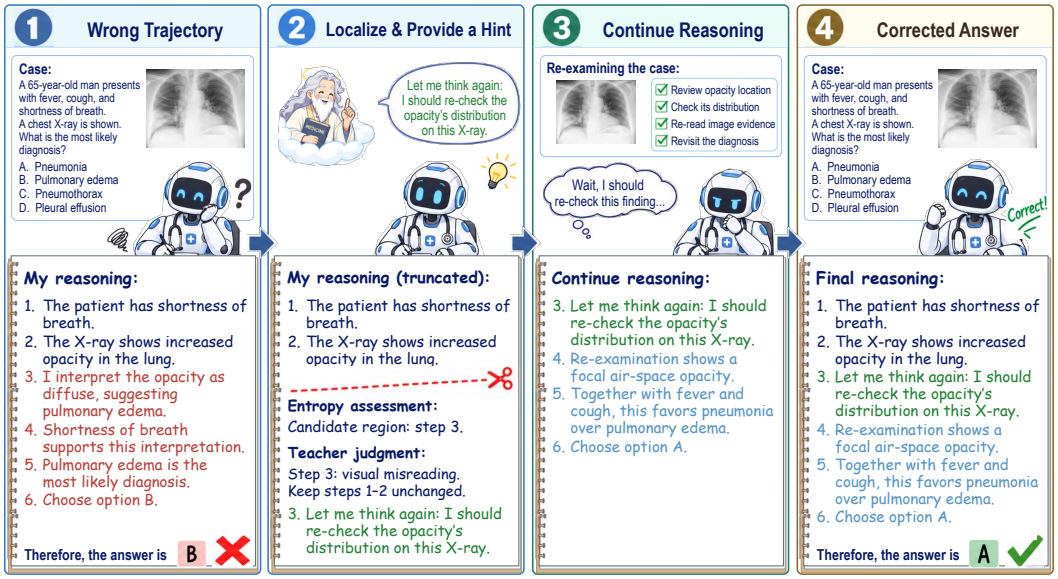


Figure 1: Local navigation from a failed reasoning trajectory. A student produces an incorrect answer; a teacher retains a supported prefix and supplies an answer-blind hint; the same student re-examines the evidence and continues to a corrected answer. The comic illustrates the interface; localization and admission checks are detailed in Section 3.

2 Related Work

2.1 Medical Reasoning, Distillation, and Data Selection

Medical vision-language models have progressed from biomedical pretraining and instruction tuning(Li et al., 2023; Moor et al., 2023) to generalist understanding and reasoning models such as Lingshu(LASA Team et al., 2025). Verifier-guided search, knowledge-enhanced synthesis, and reinforcement learning(Chen et al., 2024; Liu et al., 2025; Li et al., 2026a), together with parallel reasoning and multi-agent collaboration(Chen et al., 2026; Zhao et al., 2026b; Xia et al., 2026),

improve reasoning capabilities. Process-level diagnostics and distillation audits distinguish answer correctness from reliable visual recognition, knowledge application, and evidence integration(Thapa et al., 2025; Yang et al., 2026; Jiang et al., 2026).

Process supervision provides intermediate-step feedback(Lightman et al., 2024), while trajectory synthesis and selection address rationale quality, difficulty, and distribution(Lin et al., 2026; Chandra et al., 2025). On-policy distillation emphasizes compatibility between student states and teacher capabilities(Li et al., 2026b); reference-induced shortcuts, reasoning paths, and state coverage also shape supervision value(Shen et al., 2026; Hou et al., 2026; Fu et al., 2026). MedEvo connects medical error diagnosis with student-centered supervision, using minimal teacher hints for local correction and retaining student reasoning for self-distillation.

2.2 Learning from Failed Trajectories and Localized Repair

Failed trajectories contain reusable reasoning as well as errors. Self-Refine and Reflexion improve subsequent attempts through self-feedback and verbal memory without weight updates (Madaan et al., 2023; Shinn et al., 2023). LEMA constructs mistake–correction pairs, TrajFusion combines incorrect and correct trajectories, and EEF extracts useful actions from failed demonstrations (An et al., 2023; Deng et al., 2026; Lan et al., 2025). HHD and FailForge further guide recovery with concise hints or procedural skills rather than complete teacher-written solutions (Wang et al., 2026b; Lv et al., 2026).

Local recovery also depends on where guidance enters and how it becomes supervision. Error-localization studies, MOTAB, TTEL, and Selective Regenerative Decoding investigate localization, backtracking corrections, and prefix reuse (Tyen et al., 2024; Wang et al., 2026a; Chitale et al., 2026; Pu et al., 2026). SCoRe learns self-correction through reinforcement learning; Project Aletheia and Pyligent derive supervision from verified backtracking traces or search trees; TAPO combines micro-reflective trajectories with capability-dependent selection (Kumar et al., 2025; Dixit et al., 2026; Beresnev et al., 2026; Huang et al., 2026). EvoThink and ROSD instead use preference learning and reflective on-policy self-distillation (Dai et al., 2026; Zhao et al., 2026c).

FATE, RedFlow, and FLARE also organize recovery using structured failure feedback in agent-safety and robotics settings (Yin et al., 2026; Yan et al., 2026; Zhao et al., 2026a). Within this failure-aware learning tradition, MedEvo focuses on medical-contextualized, same-student local navigation: preserving supported prefixes, guiding continuation with gold-blind auditing and tag-conditioned hints, and selecting supervision by student-relative learnability.

2.3 Data Evolution and Recursive Self-Improvement

STaR and ReST^{EM} iteratively construct supervision through self-generated reasoning and outcome selection(Zelikman et al., 2022; Singh et al., 2023); EvoThink and ROSD further explore preference learning and reflective self-distillation(Dai et al., 2026; Zhao et al., 2026c). Difficulty-guided synthesis, multimodal generation workflows, and capability-dependent selection extend these updates to the supervision distribution(Liu et al., 2026b; Greco et al., 2026; Huang et al., 2026).

Recursive self-improvement examines how feedback changes subsequent learning experience(Duan et al., 2026), including data-strategy optimization, evolving environments, and medical curricula(Meng et al., 2026; Zheng et al., 2026; Liu et al., 2026a). MedEvo diagnoses medical groups and error patterns on a fixed feedback set, then generates and repairs corresponding training-pool trajectories to turn capability feedback into next-round self-distillation supervision.

3 Method

Overview. MedEvo has three parts: defining medical tags and trajectory errors, constructing supervision through localization and hint-guided continuation, and updating subsequent training data through feedback. Given a medical multimodal question $x_i = (I_i, q_i, O_i)$, the current student $M_t = \pi_{\theta_t}$ generates reasoning trajectories. An external auditor localizes failures and provides medical-tag-conditioned recovery hints, after which the same student regenerates the erroneous suffix. Accepted repairs are combined with original correct trajectories to construct the next training dataset.

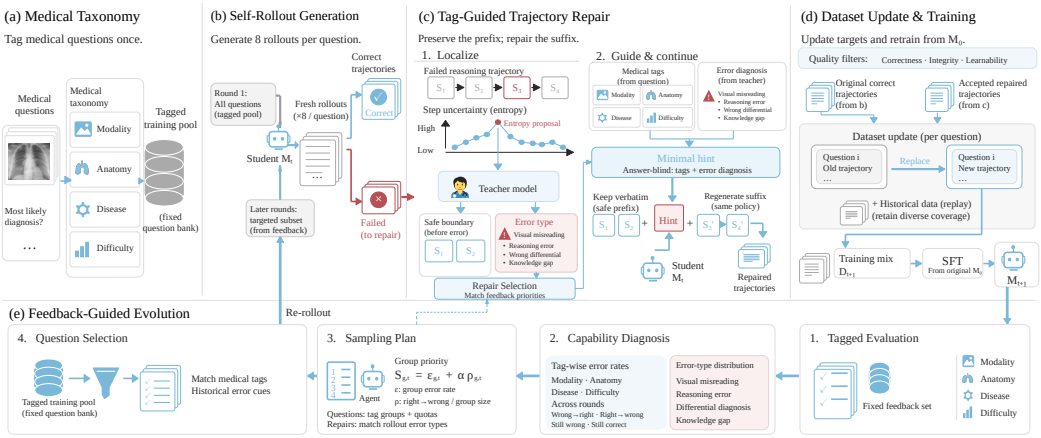


Figure 2: MedEvo overview. (a) Medical questions are tagged once by modality, anatomy, disease, and difficulty. (b) The current model generates fresh rollouts for all questions in the first round and feedback-selected questions in later rounds. (c) A gold-blind auditor localizes failures and provides tag-guided minimal hints; the same student preserves a safe prefix and regenerates the suffix. (d) Correct and accepted repaired trajectories update the dataset by question, retaining valid historical targets and replacing them only with accepted improvements, before training the next model from M_0. (e) Fixed-set evaluation, capability diagnosis, data planning, and targeted question retrieval guide the next rollout round.

We distinguish the generation policy from the training initialization. The current checkpoint M_t generates both the failed rollout and its repaired continuation, whereas every training round starts from the same original checkpoint M_0. Experience is transferred between rounds through the updated supervision dataset. The first round establishes broad modality coverage; subsequent rounds use capability and error feedback to prioritize questions for new rollouts and repairs.

The candidate training pool $\mathcal{T}$, feedback set V_{feedback}, and final test set have separate roles. The fixed feedback set guides selection and also serves as development data for calibrating selection and filtering parameters, and final test results do not influence data construction. Reference answers identify failed training rollouts and verify generated answers, but are withheld from auditing and continuation.

3.1 Medical Taxonomy and Error Types

Each question has four-dimensional tags $z_i = (z_i^{\text{mod}}, z_i^{\text{anat}}, z_i^{\text{dis}}, z_i^{\text{diff}})$ describing modality, anatomy, disease, and difficulty. Failed trajectories are classified into visual misreading, reasoning error, differential diagnosis error, or knowledge gap, with 20 subtypes in total. Question tags describe the medical context, while the current trajectory's error type determines the repair direction; together they guide feedback-driven selection and minimal-hint construction. Auditing and hints use only information observable in the images and question. The complete hierarchy, error subtypes, and annotation rules are provided in Appendices C.1 and C.1.2.

3.2 Error Localization and Hint-Guided Trajectory Repair

The current student generates eight trajectories per question. Correct trajectories enter the supervision pool; failed trajectories undergo localization, hinting, and continuation. Step entropy and local uncertainty increases identify candidate error regions. The teacher, GPT-5.6-sol, then uses the image, question, and medical tags to locate the earliest consequential error and retain a safe prefix. It supplies a minimal recovery hint conditioned on the current error type. The same student preserves the prefix and regenerates the suffix, with at most five attempts per failed trajectory. A natural transition introduces the hint, which supplies neither the reference answer nor the final diagnosis.

This limits teacher intervention to local guidance while the student generates its own corrections for self-distillation. Candidates passing correctness, boundary, integrity, and learnability checks enter the dataset update below; hints are removed and repair supervision applies only to the new suffix. Appendix A.3.1 provides localization equations and parameters, Appendices A.3.2 and A.3.3 specify auditing and continuation, and Appendix A.2.5 details filtering and training.

3.3 Feedback-Driven Recursive Self-Improvement

Local repair determines how individual trajectories become supervision; feedback-driven evolution determines which questions and errors to prioritize next. After training, the new checkpoint is evaluated on a fixed feedback set and generates fresh training-pool trajectories, which pass through the repair and selection procedure above to update supervision.

3.3.1 Training Pool, Fixed Feedback Set, and Selection

We use 81,293 training questions from public medical case datasets and a separate, fixed 5,000-question feedback set. Each round identifies high-error and regressing medical groups on the feedback set to guide question selection and generation budgets. Fresh failed trajectories matching the feedback error types receive repair priority. Statistics and allocation rules appear in Appendix A.1.4.

3.3.2 Fresh Checkpoint Rollouts and Fixed-Origin Retraining

Let D_t be the active dataset used to train M_t, with $D_0 = \varnothing$. The first round assembles original correct and accepted repaired trajectories with modality-balanced sampling. Subsequent rounds retain the existing valid supervision and update selected questions according to feedback.

For selected questions, M_t generates eight fresh rollouts per question. Original correct trajectories enter the quality-screened candidate pool directly; failed trajectories undergo the localized repair procedure. New candidates are admitted according to Appendix A.2.5. We maintain at most one active training target per canonical question, while retaining all trajectory versions and their error diagnoses in a separate history.

Retention and replacement. A historical target is replaced only when an accepted candidate has better evidence support, reasoning consistency, and conciseness; otherwise, it is retained. Each question has one active training target, with error diagnoses stored separately. For trajectories with multiple errors, repair starts before the earliest consequential error and checks the entire regenerated suffix.

Let $R_t \subseteq D_t$ contain the historical targets selected for replacement, U_{t+1} their accepted replacements, and A_{t+1} accepted targets for previously unrepresented questions. The updated dataset is

$$D_{t+1} = (D_t \setminus R_t) \cup U_{t+1} \cup A_{t+1}, \tag{1}$$

$$M_{t+1} = \mathrm{SFT}(M_0, D_{t+1}). \tag{2}$$

If the first round covers the full question pool, $A_{t+1} = \varnothing$ and later rounds only update existing targets. Each round trains on the complete updated dataset from M_0, resetting the optimizer and learning-rate schedule. The current checkpoint provides feedback and new rollouts, rather than the initialization for the next model.

4 Experiments

We use Qwen3.5-9B as the main backbone and evaluate seven medical multimodal benchmarks with MenUniEval. The training bank contains 81,293 questions from the training splits of public medical datasets, deduplicated during construction. Deduplication records include question and source IDs, available patient/case IDs, image hashes, and near-duplicate clusters (Appendix C.1.4). A fixed 5,000-question feedback set guides capability diagnosis and subsequent data construction. Each

round regenerates trajectories and retrains from the fixed initial model. The 35B and 397B experiments assess applicability across model scales.

4.1 Main Results on Medical Multimodal Benchmarks

Table 1: Accuracy (%) on seven medical benchmarks. Avg. is the arithmetic mean; rows list proprietary models, Base, method baselines, direct distillation, and MedEvo.

Model	PMC-VQA	VQA-RAD	SLAKE	PathVQA	OmniMedVQA	MedXQA	MMMU-Med	Avg.
Proprietary models								
GPT-5.6-sol	60.43	70.29	81.34	54.84	76.67	77.50	86.24	72.47
Gemini-3.8-Flash	64.52	74.72	81.34	58.58	80.76	86.30	89.10	76.47
Open-source models (< 10B)								
LLaVA-Med-7B	30.50	53.70	48.00	38.80	44.30	20.30	29.30	37.84
HuatuoGPT-V-7B	53.30	67.00	67.80	48.00	74.20	21.60	47.30	54.17
BioMediX2-8B	43.50	49.20	57.70	37.00	63.30	21.80	39.80	44.61
InternVL2.5-8B	51.30	59.40	69.00	42.10	81.30	21.70	53.50	54.04
InternVL3-8B	53.80	65.40	72.80	48.60	79.10	22.40	59.20	57.33
LINGSHU-7B	56.30	67.90	83.10	61.90	82.90	26.70	54.00	61.83
Qwen3.5-9B (Base)	59.20	68.10	76.80	48.80	82.20	47.40	76.50	65.57
Baseline 1 (HHD)	58.93	67.65	71.45	50.12	82.35	48.75	76.90	65.16
Baseline 2 (TrajFusion)	59.32	69.12	77.23	49.23	83.46	49.56	77.20	66.45
Baseline 3 (Teacher SFT)	61.23	70.15	71.60	53.20	81.90	52.30	78.14	66.93
MedEvo (9B)	**62.68**	**70.40**	**78.23**	**53.34**	**85.41**	**54.50**	**79.28**	**69.12**
Open-source models (> 10B)								
HealthGPT-14B	56.40	65.00	66.10	56.70	75.20	24.70	49.60	56.24
HuatuoGPT-V-34B	56.60	61.40	69.50	44.40	74.00	22.10	51.80	54.26
InternVL3-14B	54.10	66.30	72.80	48.00	78.90	23.10	63.10	58.04
InternVL2.5-38B	57.20	61.40	70.30	46.90	79.90	24.40	61.60	57.39
InternVL3-38B	56.60	65.40	72.70	51.00	79.80	25.20	65.20	59.41
Qwen3.5-35B (Base)	62.40	70.50	70.10	49.15	83.20	60.00	80.00	67.91
MedEvo (35B)	**64.63**	**75.83**	**80.30**	**56.01**	**83.97**	**66.50**	**84.02**	**73.04**

Paired bootstrap confidence intervals are reported in Appendix Table 7.

For 9B, the mean increases from 65.57% to 69.12%, a gain of 3.55 points. MedXpertQA-MM improves from 47.40% to 54.50%. All seven scores exceed the corresponding Base. Mean gains over HHD, TrajFusion, and teacher-distillation SFT are 3.96, 2.67, and 2.19 points, respectively. The 35B mean rises from 67.91% to 73.04%; round-wise results are reported in the scale analysis.

4.2 Round-wise Evolution of the 9B Model

Table 2: Round-wise benchmark accuracy (%) for 9B.

Round	PMC-VQA	VQA-RAD	SLAKE	PathVQA	OmniMedVQA	MedXQA	MMMU-Med	Avg.
R0	59.20	68.10	76.80	48.80	82.20	47.40	76.50	65.57
R1	59.54	68.47	76.92	52.38	80.81	52.56	79.83	67.22
R2	60.97	68.06	77.34	51.27	81.58	52.79	79.91	67.42
R3	62.68	70.40	78.23	53.34	85.41	54.50	79.28	69.12

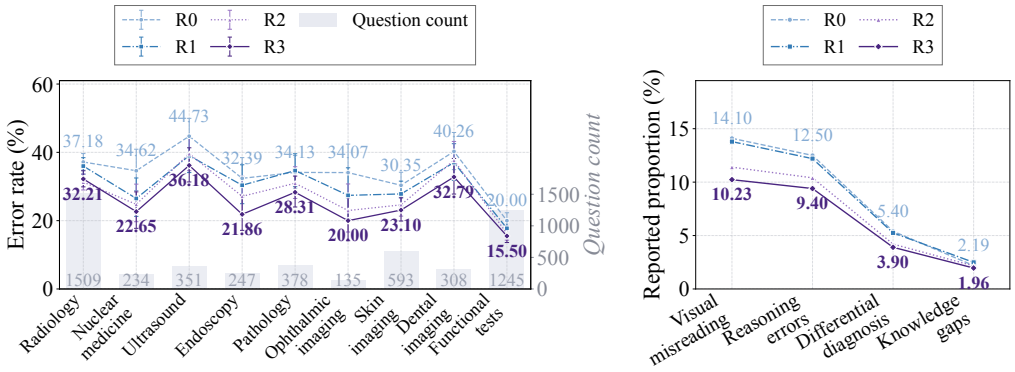
The mean progresses from 65.57% to 67.22%, 67.42%, and 69.12%. The final round exceeds Round 1 by 1.90 points, indicating an additional benefit from subsequent data updates. Individual tasks fluctuate: PathVQA scores 52.38%, 51.27%, and 53.34% across trained rounds. Round-wise construction counts and repair-volume analysis are provided in Appendix B.2.4.

4.3 Medical Capabilities and Errors on the Feedback Set

Detailed modality and error-family statistics appear in Appendix B.2.7 (Tables 11 and 12).

Feedback-driven evolution reduces errors across modalities, with stronger improvements in visual and reasoning errors than in persistent knowledge-related errors.

Figure 3 compares R0–R3 modality error rates and reported error-family proportions. All modality groups improve from R0 to R3. Visual and reasoning errors show larger absolute reductions than knowledge gaps.



(a) Modality errors (b) Error types

Figure 3: Feedback error profiles across R0–R3. (a) Modality error rates, with background bars showing group sizes on the right axis; error bars are pointwise 95% Wilson intervals under independent-question binomial sampling. (b) Reported error-family proportions. Lines connect discrete categories only to aid comparison.

Round-wise changes across twenty error subtypes are analyzed in Appendix B.2.6 (Figure 7).

5 Ablation Studies and Analysis

We examine trajectory construction, feedback across rounds, model scale, data selection, and continuation budgets. All analyses except the scale experiments use the 9B backbone.

5.1 Effectiveness of First-round Trajectory Navigation

We compare hint-based full rewriting, no-hint rewriting, generic-hint continuation, and type-guided continuation in the first 9B round, measuring repair success and post-training seven-benchmark mean accuracy.

Ours achieves 71.80% repair success and 67.22% post-training mean accuracy, exceeding generic-hint continuation by 7.30 and 3.02 points and hint-based full rewriting by 14.00 and 4.02 points. These results support type-conditioned guidance while retaining the valid student prefix.

Table 3: First-round construction controls (%). Repair success and post-training seven-benchmark mean accuracy are reported separately; subsequent rounds use Ours.

Method	Repair success	Mean accuracy
Hint + full rewrite	57.80	63.20
No-hint rewrite	36.70	61.16
Generic hint + suffix continuation	64.50	64.20
Type-guided hint + suffix continuation (Ours)	71.80	67.22

Figure 4 shows both repair success and downstream accuracy, on which the full method exceeds the compared configurations.

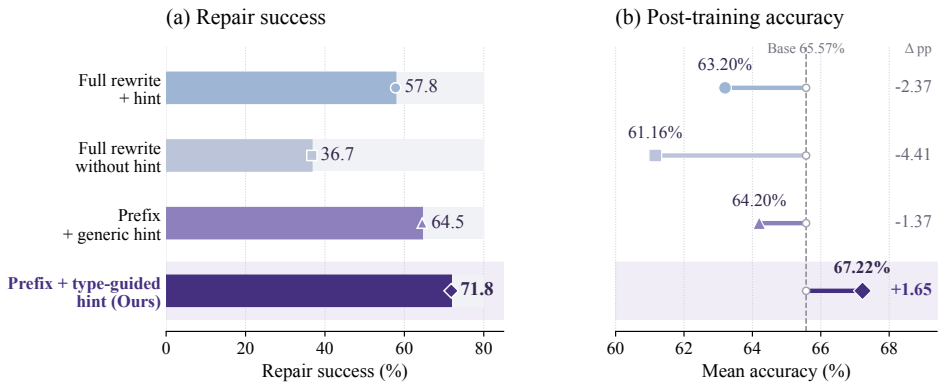


Figure 4: First-round trajectory-navigation comparisons. (a) Repair success for four configurations. (b) Post-training seven-benchmark mean; the dashed line marks Base at 65.57%.

5.2 Contributions of Feedback to Multi-round Evolution

We compare no-feedback, no-medical-tag, no-error-type-feedback, and full MedEvo runs to assess the feedback system. All share R1 (67.22%); we compare feedback configurations over the following two rounds.

Full MedEvo improves from the shared R1 accuracy of 67.22% to 69.12% at R3, a gain of 1.90 points. It finishes 2.45, 1.59, and 2.91 points above the repair-only, no-medical-tag, and no-error-type-feedback variants, respectively.

Table 4: Round-wise feedback comparisons: seven-benchmark mean accuracy (%). All variants share R1; the last column reports R3 minus R1 in percentage points.

Method	R1	R2	R3	Δ
No feedback (repair only)	67.22	66.39	66.67	−0.55
Without medical tags	67.22	67.32	67.53	+0.31
Without error-type feedback	67.22	67.01	66.21	−1.01
Full MedEvo	**67.22**	**67.42**	**69.12**	**+1.90**

5.3 Generalization across Model Scales

We apply the repair and feedback-driven evolution procedure to larger backbones to test whether its benefits extend beyond 9B. We compare Base with three training rounds for 35B and two for 397B, reporting individual accuracies and means on the same seven benchmarks.

5.3.1 Multi-round Experiments on 35B

Table 5: Round-wise benchmark accuracy (%) for 35B.

Round	PMC-VQA	VQA-RAD	SLAKE	PathVQA	OmniMedVQA	MedXQA	MMMU-Med	Avg.
R0	62.40	70.50	70.10	49.15	83.20	60.00	80.00	67.91
R1	63.12	62.08	78.14	52.71	85.39	65.68	81.02	69.73
R2	60.33	70.91	79.18	50.87	85.22	66.73	82.41	70.81
R3	64.63	75.83	80.30	56.01	83.97	66.50	84.02	73.04

The 35B mean rises from 67.91% to 69.73%, 70.81%, and 73.04%, a final gain of 5.13 points. MedXpertQA-MM peaks at 66.73% in Round 2 and falls to 66.50% in Round 3, showing task-dependent optima.

5.3.2 Two-round Validation on 397B

Given training-time and resource constraints, the 397B evaluation includes two training rounds.

Table 6: Benchmark accuracy (%) for 397B.

Round	PMC-VQA	VQA-RAD	SLAKE	PathVQA	OmniMedVQA	MedXQA	MMMU-Med	Avg.
R0	63.47	74.72	77.38	52.49	85.10	67.65	82.03	71.83
R1	64.87	76.05	78.89	58.51	85.15	74.40	85.10	74.71
R2	65.72	77.15	79.10	57.90	85.05	75.00	84.75	74.95

The 397B mean increases from 71.83% to 74.71% and 74.95%, a Round-2 gain of 3.12 points over Base; MedXpertQA-MM rises from 67.65% to 75.00%. Six of seven Round-2 benchmarks exceed Base; OmniMedVQA decreases from 85.10% to 85.05% (0.05 percentage points). These observations support applicability across scales.

Mean accuracy increases across reported rounds for 35B and 397B, while individual benchmarks may regress.

Discussion

MedEvo connects the correction of a reasoning trajectory to the choice of a future training example. Retaining the student prefix limits how much valid reasoning is discarded, while medical-tag feedback makes the curriculum responsive to the model's changing capability profile. The experiments evaluate trajectory construction and subsequent feedback-driven updates.

6 Conclusion

Its central advantage is minimal teacher guidance coupled with self-distillation: the teacher supplies a localized recovery cue, the student generates the correction, and accepted student trajectories become training supervision. This preserves useful student reasoning while turning its own corrections into reusable learning signals.

We introduced MedEvo, a framework for evolving medical multimodal reasoning through local trajectory repair and feedback-driven data refinement. MedEvo preserves valid prefixes from failed student trajectories and combines uncertainty-guided auditing with medical-tag-conditioned hints to elicit corrected continuations. Repair-aware filtering selects supervision for training alongside correct self-generated trajectories. After each round, medical capability profiles and error-type feedback guide the selection and regeneration of training data. Experiments across model scales demonstrate the effectiveness of this approach, while feedback-guided updates yield further improvements beyond the first round. Our findings suggest that the value of self-generated data depends on how well supervision addresses the current model's failures. By connecting localized correction with adaptive data construction, MedEvo offers a practical foundation for further investigation of recursive self-improvement in medical reasoning.

Declaration of AI Assistance. AI tools were used solely for language editing, manuscript formatting, and figure preparation. They were not involved in method development or experimental design. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final manuscript.

Reproducibility Statement. The method and experimental procedures are described in the main text and appendices. We will release the implementation, prompts, training configurations, evalua-

tion scripts, and data-processing instructions to support reproducibility, subject to the licenses and access restrictions of the underlying datasets.

References

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. Learning From Mistakes Makes LLM Better Reasoner. *arXiv preprint arXiv:2310.20689*, 2023. URL https://arxiv.org/abs/2310.20689.

Dmitry Beresnev, Vladimir Makharev, Roman Khalikov, Ivan V. Oseledets, and Petr Anokhin. Search, fail, recover: A training framework for correction-aware reasoning. *CoRR*, abs/2607.07492, 2026. doi: 10.48550/ARXIV.2607.07492. URL https://doi.org/10.48550/arXiv.2607.07492.

Abhranil Chandra, Ayush Agrawal, Arian Hosseini, Sebastian Fischmeister, Rishabh Agarwal, Navin Goyal, and Aaron Courville. Shape of thought: When distribution matters more than correctness in reasoning tasks. *arXiv preprint arXiv:2512.22255*, 2025.

Jianwen Chen, Xinyu Yang, Peng Xia, Arian Azarang, Yueh Z. Lee, Gang Li, Hongtu Zhu, Yun Li, Beidi Chen, and Huaxiu Yao. Medverse: Efficient and reliable medical reasoning via dag-structured parallel execution. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 15320–15336. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.699. URL https://doi.org/10.18653/v1/2026.acl-long.699.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs. *arXiv preprint arXiv:2412.18925*, 2024. URL https://arxiv.org/abs/2412.18925.

Rajiv Shailesh Chitale, Rahul Madhavan, Taneesh Gupta, Deepanway Ghosal, and Aravindan Raghuveer. Test-time scaling via error localization. *CoRR*, abs/2607.21453, 2026. doi: 10.48550/ARXIV.2607.21453. URL https://doi.org/10.48550/arXiv.2607.21453.

Xinbang Dai, Zheyu Xin, Huikang Hu, Lin Ren, Rihui Jin, Guohui Xiao, Kuicai Dong, Zhaocheng Du, and Yuyang Zhang. Evothink: Evolving thinking in large reasoning models via self-pruning and aha-moment preference optimization. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence and the 29th European Conference on Artificial Intelligence, IJCAI/ECAI 2026, Bremen, Germany, 15-21 August 2026*, pp. 5693–5701. ijcai.org, 2026. doi: 10.24963/IJCAI.2026/634. URL https://doi.org/10.24963/ijcai.2026/634.

Jie Deng, Hanshuang Tong, Jun Li, Shining Liang, Ning Wu, Hongzhi Li, and Yutao Xie. Beyond rejection sampling: Trajectory fusion for scaling mathematical reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 7943–7959. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.390. URL https://doi.org/10.18653/v1/2026.findings-acl.390.

Aradhya Dixit, Tianxi Liang, and Jai Telang. Project aletheia: Verifier-guided distillation of backtracking for small language models. *arXiv preprint arXiv:2601.14290*, 2026.

Yi Duan, Ying Liu, Zirui Tang, Haodong Chen, Jun Zhou, Yumou Liu, et al. The Last AI Built by Humans: Toward Genuine Recursive Self-Improvement. *arXiv preprint arXiv:2609.11873*, 2026. URL https://arxiv.org/abs/2609.11873.

Zixuan Fu et al. Rethinking On-Policy Distillation of Large Language Models II: One Training Example. *arXiv preprint arXiv:2609.04172*, 2026. URL https://arxiv.org/abs/2609.04172.

Johnny Greco et al. NeMo Data Designer: An Extensible Framework for Multimodal Synthetic Data Generation. *arXiv preprint arXiv:2609.17699*, 2026. URL https://arxiv.org/abs/2609.17699.

Zhinan Hou et al. What Matters in On-Policy Distillation? A Perspective on Data Efficiency and Data Selection. *arXiv preprint arXiv:2609.05198*, 2026. URL https://arxiv.org/abs/2609.05198.

Zhilin Huang, Hang Gao, Ziqiang Dong, Yuan Chen, Yifeng Luo, Chujun Qin, Jingyi Wang, Yang Yang, and Guanjun Jiang. Learning from your own mistakes: Constructing learnable micro-reflective trajectories for self-distillation. *CoRR*, abs/2606.18844, 2026. doi: 10.48550/ARXIV.2606.18844. URL https://doi.org/10.48550/arXiv.2606.18844.

Zhaoyang Jiang, Xuanqi Peng, Fei Teng, Zhizhong Fu, Yunsoo Kim, Jiacong Mi, Zicheng Li, and Honghan Wu. Better accuracies, worse reasoning: A step-level audit of medical chain-of-thought distillation. *CoRR*, abs/2605.28301, 2026. doi: 10.48550/ARXIV.2605.28301. URL https://doi.org/10.48550/arXiv.2605.28301.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D. Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M. Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal M. P. Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=CjwERcAU7w.

Li-Cheng Lan, Andrew Bai, Minhao Cheng, Cho-Jui Hsieh, and Tianyi Zhou. Exploring Expert Failures Improves LLM Agent Tuning. *arXiv preprint arXiv:2504.13145*, 2025. URL https://arxiv.org/abs/2504.13145.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, et al. Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning. *arXiv preprint arXiv:2506.07044*, 2025. URL https://arxiv.org/abs/2506.07044.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *arXiv preprint arXiv:2306.00890*, 2023. URL https://arxiv.org/abs/2306.00890.

Haolin Li, Shuyang Jiang, Ruipeng Zhang, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Eliciting medical reasoning with knowledge-enhanced data synthesis: A semi-supervised reinforcement learning approach. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 2847–2870. Association for Computational Linguistics, 2026a. doi: 10.18653/V1/2026.FINDINGS-ACL.137. URL https://doi.org/10.18653/v1/2026.findings-acl.137.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Proceedings of NAACL-HLT*, 2024. URL https://aclanthology.org/2024.naacl-long.421/.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Honglin Lin, Zheng Liu, Yun Zhu, Chonghan Qin, Juekai Lin, Xiaoran Shang, Conghui He, Wentao Zhang, and Lijun Wu. Mmfinereason: Closing the multimodal reasoning gap via open data-centric methods. *CoRR*, abs/2601.21821, 2026. doi: 10.48550/ARXIV.2601.21821. URL https://doi.org/10.48550/arXiv.2601.21821.

Chi Liu, Derek Li, Yan Shu, Robin Chen, Derek Duan, Teng Fang, and Bryan Dai. Fleming-r1: Toward expert-level medical reasoning via reinforcement learning. *CoRR*, abs/2509.15279, 2025. doi: 10.48550/ARXIV.2509.15279. URL https://doi.org/10.48550/arXiv.2509.15279.

Junqi Liu, Yufan He, Yexiao He, Pengfei Guo, Dong Yang, Andriy Myronenko, Can Zhao, Hanrong Ye, Tianhao Qi, Yuyin Zhou, Daguang Xu, and Yucheng Tang. Bat: Towards self-evolving medical research agent with stage rubrics. *CoRR*, abs/2608.16211, 2026a. doi: 10.48550/ARXIV.2608.16211. URL https://doi.org/10.48550/arXiv.2608.16211.

Zheng Liu, Honglin Lin, Xiaoyang Wang, Xin Gao, Yu Li, Mengzhang Cai, Yun Zhu, Zhanping Zhong, Qizhi Pei, Zhuoshi Pan, Xiaoran Shang, Conghui He, Bin Cui, Wentao Zhang, and Lijun Wu. Chartverse: Scaling chart reasoning via reliable programmatic synthesis from scratch. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 7551–7577. Association for Computational Linguistics, 2026b. doi: 10.18653/V1/2026.ACL-LONG.344. URL https://doi.org/10.18653/v1/2026.acl-long.344.

Dongyi Lv et al. FailForge: Distilling Procedural Competence from Persistent Failures into Code Agents. *arXiv preprint arXiv:2608.08570*, 2026. URL https://arxiv.org/abs/2608.08570.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

Fanqing Meng, Lingxiao Du, Qiguang Chen, Ziqi Zhao, Haocheng Lu, Mengkang Hu, and Michael Qizhe Shieh. Rsibench-data: Benchmarking data-centric research for recursive self-improvement. *CoRR*, abs/2607.25886, 2026. doi: 10.48550/ARXIV.2607.25886. URL https://doi.org/10.48550/arXiv.2607.25886.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, et al. Med-Flamingo: a Multimodal Medical Few-shot Learner. *arXiv preprint arXiv:2307.15189*, 2023. URL https://arxiv.org/abs/2307.15189.

Sophia Xiao Pu, Yumo Xu, Sailik Sengupta, Millennium Bismay, Ruixue Lian, James Gung, Yi-An Lai, and Arshit Gupta. Selective regenerative decoding: Trajectory-level intervention for inference-time reasoning. *CoRR*, abs/2608.24338, 2026. doi: 10.48550/ARXIV.2608.24338. URL https://doi.org/10.48550/arXiv.2608.24338.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, Gang Chen, and Jieping Ye. Purified OPSD: on-policy self-distillation without losing how to think. *CoRR*, abs/2607.02234, 2026. doi: 10.48550/ARXIV.2607.02234. URL https://doi.org/10.48550/arXiv.2607.02234.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023. URL https://arxiv.org/abs/2312.06585.

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, Shriya Reddy, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. Disentangling reasoning and knowledge in medical large language models. *CoRR*, abs/2505.11462, 2025. doi: 10.48550/ARXIV.2505.11462. URL https://doi.org/10.48550/arXiv.2505.11462.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. Llms cannot find reasoning errors, but can correct them given the error location. *Findings of ACL*, 2024.

Bing Wang, Shaotian Yan, Chen Shen, Kaiyuan Liu, Sinan Fan, Ximing Li, Rui Miao, Xiaosong Yuan, Zhanming Shen, and Jieping Ye. Backtracking when it strays: Mitigating dual exposure biases in LLM reasoning distillation. *CoRR*, abs/2605.19433, 2026a. doi: 10.48550/ARXIV.2605.19433. URL https://doi.org/10.48550/arXiv.2605.19433.

Shengjie Wang, Guanghe Li, Zonghan Yang, and Yang Gao. Hindsight hint distillation: Scaffolded reasoning for SWE agents from cot-free answers. *CoRR*, abs/2605.11556, 2026b. doi: 10.48550/ARXIV.2605.11556. URL https://doi.org/10.48550/arXiv.2605.11556.

Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Zihan Dong, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Linjun Zhang, Shujie Liu, Yan Lu, and Huaxiu Yao. MMedAgent-RL: Optimizing Multi-Agent Collaboration for Multimodal Medical Reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/83e25d16305255ae61e2cfb34639e177-Abstract-Conference.html.

Zhengyang Yan et al. RedFlow: Redirect Failure into Action-Level Corrections for Flow-matching VLA Policy. *arXiv preprint arXiv:2607.27782*, 2026. URL https://arxiv.org/abs/2607.27782.

Sicheng Yang, Hangjie Yuan, Wenjun Zhang, Jinwang Wang, Yichen Qian, Weihua Chen, Fan Wang, and Lei Zhu. Clinhallu: A benchmark for diagnosing stage-wise hallucinations in medical MLLM reasoning. *CoRR*, abs/2606.14697, 2026. doi: 10.48550/ARXIV.2606.14697. URL https://doi.org/10.48550/arXiv.2606.14697.

Bo Yin, Qi Li, and Xinchao Wang. On-Policy Self-Evolution via Failure Trajectories for Agentic Safety Alignment. *arXiv preprint arXiv:2605.11882*, 2026. URL https://arxiv.org/abs/2605.11882.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Self-taught reasoner—bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ganlong Zhao et al. FLARE: A Failure-Aware Framework for Autonomous Correction and Recovery in Visual-Language Robotic Manipulation. *arXiv preprint arXiv:2608.26645*, 2026a. URL https://arxiv.org/abs/2608.26645.

Yu Zhao, Hao Guan, Yongcheng Jing, Ying Zhang, and Dacheng Tao. Medcog: Maximizing LLM inference density in medical reasoning via meta-cognitive regulation. *CoRR*, abs/2602.07905, 2026b. doi: 10.48550/ARXIV.2602.07905. URL https://doi.org/10.48550/arXiv.2602.07905.

Ziqi Zhao, Xinyu Ma, Liu Yang, Yujie Feng, Daiting Shi, Jingzhou He, Xin Xin, Zhaochun Ren, and Xiao-Ming Wu. ROSD: reflective on-policy self-distillation for language model reasoning across domains. *CoRR*, abs/2605.28014, 2026c. doi: 10.48550/ARXIV.2605.28014. URL https://doi.org/10.48550/arXiv.2605.28014.

Tong Zheng et al. Dream-RSI: Recursive Self-Improvement through Evolving Worlds. *arXiv preprint arXiv:2609.14858*, 2026. URL https://arxiv.org/abs/2609.14858.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/zuo25a.html.

A Method Details

A.1 Navigation and Evolution Implementation Details

A.1.1 Entropy Proposal and Window Construction

For token j in a failed rollout, compute $H_j = -\sum_v p_{\theta_t}(v \mid x, r^-_{<j}) \log p_{\theta_t}(v \mid x, r^-_{<j})$. The semantic-step score and its local increase are

$$E^*_k = \lambda \,\mathrm{Mean}_{j \in s_k} H_j + (1-\lambda)\,\mathrm{P90}_{j \in s_k} H_j, \qquad \Delta E_k = E^*_k - \frac{1}{m_k} \sum_{j=k-m_k}^{k-1} E^*_j. \tag{3}$$

For $k > 1$, $m_k = \min(m, k-1)$; set $\Delta E_1 = 0$. Normalize each profile, using zero for a constant profile, then choose

$$k^* = \arg\max_k [\widetilde{E}^*_k + \widetilde{\Delta E}_k], \qquad \mathcal{W}(k^*) = \{s_{k^*-w_\ell}, \ldots, s_{k^*+w_r}\}. \tag{4}$$

Clip the window to valid step indices. The segmenter, normalization, m, λ, and window widths are fixed in the run configuration. The auditor may move outside this proposal when tracing an unsupported premise.

A.1.2 Calibration and Training Targets

The actual configuration uses $\lambda = 0.5$, $m = 3$, Min-Max normalization, and window widths $w_\ell = w_r = 1$. Learnability thresholds are $D_{\text{repair}} \in [0.35, 0.85]$ and $\Delta_{\text{repair}} \in [0.02, 0.30]$, calibrated on the fixed 5,000-question feedback set (configuration v1.2). Using the suffix scores in Appendix A.2.4, set

$$L_i = \mathbf{1}\{D_{\min} \le D^{(i)}_{\text{repair}} \le D_{\max} \wedge \Delta_{\min} \le \Delta^{(i)}_{\text{repair}} \le \Delta_{\max}\}. \tag{5}$$

The thresholds above are calibrated on the fixed feedback set and frozen before final evaluation. Record the values, calibration split, retained fraction, and generator hash. Score the suffix under the current generator without the construction hint.

Repaired examples mask the retained prefix and use

$$\mathcal{L}_{\text{suffix}} = -\sum_{j=\ell_b+1}^{|r^*|} \log \pi_\theta(r^*_j \mid x, r^*_{<j}), \tag{6}$$

where ℓ_b is the prefix token count. Original correct examples use full-response loss. Remove the construction hint before SFT; retained historical repairs keep their stored boundaries and masks. No external hint is supplied at evaluation. Record the collator and masks with each run; loss masking and suffix-level data filtering are separate operations.

A.1.3 Group Support, Quotas, and Replay

Use supported single-axis groups and predeclared intersections. If n_g is below a fixed support threshold, back off to a supported ancestor; merge identical resolved nodes. At the initial checkpoint, inter-round transitions are unavailable and priorities use error rate alone. For later rounds, $\rho_{g,t} = N^{10}_{g,t}/n_g$; the conditional regression rate can also be reported with the previously correct population as denominator.

Allocate the attempt budget proportionally to group scores with largest-remainder rounding, deterministic ties, and seeded sampling. Cap quotas by available unique training questions and redistribute remaining capacity. Assign each multi-tag question once for budget accounting. Zero total scores trigger uniform sampling over eligible unique questions. Token and auditor-call caps are additional limits, and unfilled quotas remain explicit.

Deduplicate canonical questions before updating, keeping at most one active target per question. Preserve valid historical targets unless an admitted candidate improves evidence support, reasoning consistency, and unnecessary repetition under the fixed comparison rubric. Hide round and model labels from the comparison auditor; retain the historical target on ties. Preserve all versions and rollout-specific diagnoses in a separate history. New question admissions may expand coverage, but a changed error category does not create another active slot. Retrain from M_0 on the complete updated dataset, with stored suffix masks for historical repairs. Generation caps are fixed in advance. Matched selection and initialization ablations additionally hold their prescribed data and token budgets fixed. Planner records and information permissions follow Appendix C.1.4.

A.1.4 Training Pool, Fixed Feedback Set, and Selection

Fixed feedback set and test isolation. Our training data comprise 81,293 questions drawn from cases and their associated images and question–answer pairs in public medical datasets. We use a separate, fixed feedback set of 5,000 questions stratified by modality to cover medical modalities. The feedback set supports capability assessment and data selection.

After each training round, M_t is evaluated on the fixed tagged feedback set. For a supported tag group g, let $c_{i,t} \in \{0, 1\}$ indicate answer correctness. We compute

$$n_g = |V_{\text{feedback},g}|, \qquad \epsilon_{g,t} = \frac{1}{n_g} \sum_{i \in g} (1 - c_{i,t}), \tag{7}$$

$$N_{g,t}^{ab} = \sum_{i \in g} \mathbf{1}[c_{i,t-1} = a,\ c_{i,t} = b]. \tag{8}$$

The transitions distinguish corrected errors (N^{01}), regressions (N^{10}), persistent errors (N^{00}), and retained successes (N^{11}). Failed feedback responses are also diagnosed to identify the current error distribution within each group. Each question is counted once within a group, and overall statistics count the union of questions rather than summing overlapping groups. Sparse groups are aggregated to supported parent nodes.

The planner combines these statistics with candidate-pool availability to allocate the next generation budget. A reproducible rule baseline ranks groups by

$$S_{g,t} = \epsilon_{g,t} + \alpha \rho_{g,t}, \qquad \rho_{g,t} = \frac{N_{g,t}^{10}}{n_g}, \tag{9}$$

where $\alpha \geq 0$ controls the emphasis on regressions. Group quotas are proportional to these scores, capped by available unique questions, with unused quota redistributed among eligible groups. Multi-tag questions consume the budget only once; deterministic quota rounding, zero-score fallback, and support thresholds follow Appendix A.1.3. A model-based planner can additionally use the diagnosed error patterns to prioritize candidates under the same budget.

Retrieval first matches medical tags and then uses historical error records, when available, as prioritization evidence. Historical errors guide retrieval; selected questions receive fresh rollouts from M_t, and their actual errors are diagnosed again. Candidates may include previously trained questions and questions not yet admitted to training.

Among fresh rollouts of selected questions, failures matching the feedback-identified error types receive repair priority. Matching uses current diagnoses rather than inherited historical labels. This priority allocates repair effort and does not replace correctness or quality screening.

A.2 Detailed Inner Repair Formulation

A.2.1 Failure-Conditioned Trajectory Repair

In a round, π_θ denotes the frozen generator M_t. Let a medical multimodal question be $x = (I, q, O)$, where I contains one or more medical images, q is the clinical question and context, and O is the answer-choice set. Given x, a target policy π_θ generates a textual reasoning trajectory $r = (r_1, \ldots, r_T)$ and an answer $\hat{y}$:

$$(r, \hat{y}) \sim \pi_\theta(\cdot \mid x). \tag{10}$$

The reference answer is $y \in O$. We focus on an initially failed trajectory r^- for which $\hat{y} \neq y$ or no valid answer can be parsed.

Failed trajectories can retain reusable reasoning steps. We segment it into K semantically complete reasoning steps $r^- = (s_1, \ldots, s_K)$ (image observation, key feature, mechanism, differential, option mapping) and assume it contains a reusable prefix followed by an error-bearing suffix:

$$r^- = p_b \oplus s^-, \qquad p_b = (s_1, \ldots, s_b), \qquad s^- = (s_{b+1}, \ldots, s_K). \tag{11}$$

The boundary b is the latest step boundary preceding the rewrite region. The prefix p_b contains the retained steps before the diagnosed failure.

We introduce an external auditor $\mathcal{T}$ that uses the entropy proposal $\mathcal{E}(r)$ as a localization prior, diagnoses the first consequential failure in the student trajectory, and emits an *operation-oriented*, answer-blind hint rather than a corrected solution:

$$(e_t, b_t, h_t) = \mathcal{T}(q, I, O, r; \mathcal{E}(r), z^{\text{obs}}), \tag{12}$$

where e_t is the first consequential failure, b_t is the safe-prefix boundary, h_t is the minimal hint passed to the student, and $\mathcal{E}(r)$ contains the stepwise entropy values, normalized scores, local surges, proposed anchor, and candidate window defined below. The auditor is gold-blind: h_t must not reveal an answer, a final diagnosis, or a teacher-authored chain of thought. The auditor may retain a high-entropy but valid step or move the boundary earlier when a later surge is caused by an upstream unsupported premise.

The hint type is drawn from four intervention operators:

$$h_t \in \{\mathcal{P}_{\text{inspect}}, \mathcal{R}_{\text{reconstruct}}, \mathcal{D}_{\text{disambiguate}}, \mathcal{K}_{\text{bridge}}\}, \tag{13}$$

corresponding to perceptual re-examination, inference reconstruction, differential disambiguation, and knowledge bridging. These are not error labels; they specify the next missing cognitive operation required for recovery.

Given the safe prefix and hint, the repair operation constructs a new suffix

$$s^+ \sim \pi_\theta(\cdot \mid x, p_b, g(h_t)), \tag{14}$$

where $g(h_t)$ renders the recovery operation as a first-person cue. The user message contains x, and the unfinished assistant message contains $p_b \oplus g(h_t)$. The target policy continues this same message; the structured audit is retained as metadata. The repaired trajectory is $r^* = p_b \oplus s^+$, with tokens $y^*_{1:T} = p_b \oplus y^*_{\ell_b+1:T}$, where ℓ_b is the prefix token length, distinct from step index b. The cue is removed in the SFT export; only suffix tokens receive loss.

A.2.2 Entropy-Centered Audit Localization

We identify a compact rewrite region around the consequential failure. We score each step using the target policy's own uncertainty and use the dominant high-entropy region as an audit anchor.

Step-level predictive uncertainty. At step s_k the policy emits tokens $v \in \mathcal{V}$ with $p_\theta(v \mid x, y_{<t})$; we measure the local decision uncertainty by a step window that combines mean token entropy and its upper quantile to suppress single-token noise:

$$E^*_k = \lambda \ \text{Mean}_{t \in s_k} H_t + (1-\lambda) \ \text{P90}_{t \in s_k} H_t, \qquad H_t = -\sum_{v \in \mathcal{V}} p_\theta(v \mid x, y_{<t}) \log p_\theta(v \mid x, y_{<t}), \tag{15}$$

with $\lambda = 0.5$, $m = 3$, Min-Max normalization, and $w_\ell = w_r = 1$, and the relative entropy surge

$$\Delta E_k = E^*_k - \frac{1}{m_k} \sum_{j=k-m_k}^{k-1} E^*_j, \qquad m_k = \min(m, k-1) \quad (k > 1). \tag{16}$$

Set $\Delta E_1 = 0$; the lookback m controls the preceding context window. Clip all step indices to $1, \ldots, K$.

Audit anchor and rewrite window. Let

$$k^* = \arg\max_k [\widetilde{E}^*_k + \widetilde{\Delta E}_k] \tag{17}$$

denote the step with the strongest normalized uncertainty signal inside the trajectory. We then define a compact audit window

$$\mathcal{W}(k^*) = \{s_{k^*-w_\ell}, \ldots, s_{k^*+w_r}\}, \tag{18}$$

with a small fixed local context (w_ℓ, w_r) so that the audit can prioritize the unstable region without searching the full trajectory uniformly. The last step before this window, $b_0 = k^* - w_\ell - 1$ (clipped to the valid range), is only a provisional rollback boundary. The external auditor then traces the evidence dependencies backward and sets the final boundary b at the last fully valid step before the first consequential failure; b may therefore differ from b_0 when medically or logically justified.

Operational role of entropy. Entropy proposes a compact region for auditing. A high-entropy region marks where the target policy is least stable under its own decoding dynamics, which is where auditor attention is concentrated first. The auditor explicitly classifies the proposal as HIT, PARTIAL_HIT, MISS, or NON_ERROR_UNCERTAINTY before determining the evidence-grounded boundary.

A.2.3 Operation-Oriented Teacher Hints

The external teacher provides local guidance to compress diagnosis into the smallest hint that can restart the missing operation from the preserved prefix.

Perceptual re-examination. When the student misreads the image, ECG, pathology slide, or another visual cue, $\mathcal{P}_{\text{inspect}}$ specifies what visual region or attribute must be revisited, such as lesion distribution, boundary shape, waveform morphology, stain localization, or cross-view consistency. The hint points to *what to re-check*, not the correct visual conclusion.

Inference reconstruction. When the student has assembled mostly correct evidence but breaks the evidence-to-decision chain, $\mathcal{R}_{\text{reconstruct}}$ asks it to rebuild the mapping Evidence $\rightarrow$ Mechanism $\rightarrow$ Diagnosis $\rightarrow$ Requested Decision. This targets failures such as supportive-versus-causal treatment confusion, mechanism mismatch, or incorrect management ordering.

Differential disambiguation. When the student narrows the case to plausible candidates but selects the wrong one, $\mathcal{D}_{\text{disambiguate}}$ requests comparison along the most discriminative feature dimension, such as contrast pattern, onset profile, lesion location, signature pathology, or an exclusion criterion. The hint does not reopen the entire case; it focuses the student on the decisive comparison.

Knowledge bridging. When recovery requires a missing medical concept rather than better inspection or better chaining, $\mathcal{K}_{\text{bridge}}$ provides a narrowly scoped knowledge scaffold, e.g. a diagnostic criterion, drug mechanism, disease-image association, immunohistochemical localization rule, or treatment principle. It may supply the missing concept, but it must not directly map that concept to the correct option.

A.2.4 Repair-Aware Learnability

Final-answer correctness is necessary but not sufficient for a useful repair: a repaired trajectory is informative only when it exposes a genuine student error preference and introduces a locally learnable correction. We therefore score *the repair suffix only*, since the safe prefix p_b was already generated by π_θ. Suffix scoring isolates the correction rather than letting prefix likelihood dominate the statistic.

Repair-suffix difficulty. The preserved design follows the instruction-following difficulty principle (Li et al., 2024) but restricted to the repair suffix:

$$D_{\text{repair}} = \frac{\mathcal{L}(y^*_{\ell_b+1:T} \mid x, p_b)}{\mathcal{L}(y^*_{\ell_b+1:T} \mid p_b)}, \tag{19}$$

Here $\mathcal{L}$ is summed suffix negative log-likelihood under the current generator, without the construction cue; the same suffix is used in numerator and denominator. Reject empty suffixes or a zero denominator. The numerator is the student's cost of predicting the repair given the image, question, and safe prefix, and the denominator is its cost without task evidence. Thus D_{repair} measures the repair's *task dependence*. We retain an empirically calibrated interior interval and report the two cutoffs, calibration split, and retained fraction with the final run configuration. This avoids conflating a context-dependent repair with a difficult-but-unlearnable one. This is the first coordinate evaluated by the interval gate in Equation 5.

Repair preference gap. Because the original and repaired suffixes can have different lengths, use s^- and s^+ directly and compare token-average losses:

$$\Delta_{\text{repair}} = \frac{\mathcal{L}(s^+ \mid x, p_b)}{|s^+|} - \frac{\mathcal{L}(s^- \mid x, p_b)}{|s^-|}. \tag{20}$$

A positive gap indicates that the current model assigns lower average loss to its wrong suffix than to the repair. The gate retains a calibrated interval of positive gaps. Empty suffixes are rejected before scoring. It is the second coordinate evaluated by the interval gate in Equation 5.

A.2.5 Trajectory Selection and Correct-Rollout Fallback

We score suffix difficulty D_{repair} and repair preference Δ_{repair} under M_t without the hint (Appendix A.2.4) and filter repairs using:

$$L_i = \mathbf{1}[D_{\min} \le D_{\text{repair}}^{(i)} \le D_{\max} \ \wedge \ \Delta_{\min} \le \Delta_{\text{repair}}^{(i)} \le \Delta_{\max}]. \tag{21}$$

The thresholds are $D_{\text{repair}} \in [0.35, 0.85]$ and $\Delta_{\text{repair}} \in [0.02, 0.30]$, calibrated on the fixed 5,000-question feedback set. Admission combines answer, boundary, integrity, and learnability checks:

$$\text{Accept}(i) = C_i \wedge A_i \wedge N_i \wedge L_i, \tag{22}$$

Here C_i checks answer correctness, A_i validates the prefix and boundary, and N_i checks content integrity and invalid loops. Before SFT, we remove the hint and export $r^* = p_b \oplus s^+$. Repaired examples receive loss only on the new suffix after the prefix of length ℓ_b:

$$\mathcal{L}_{\text{repair}}(\theta) = -\sum_{j=\ell_b+1}^{|r^*|} \log \pi_\theta(r_j^* \mid x, r_{<j}^*). \tag{23}$$

Original correct trajectories use full-response loss; historical repairs retain their suffix masks. Each round retrains from M_0. If repair fails, we fall back to a qualified original correct trajectory or an existing valid target; otherwise, the question is skipped.

A.2.6 Operationalizing Deep Reasoning

We use *deep reasoning* as an operational term for four observable repair behaviors: failure localization, evidence re-inspection, hypothesis revision, and controlled convergence. In our framework, the entropy-centered audit window operationalizes failure localization and evidence re-inspection, while the repaired suffix operationalizes hypothesis revision and controlled convergence. The four hint operators further clarify that the teacher controls *which recovery operation* is missing, while the student still performs the actual reasoning continuation.

A.3 Implementation Details of Local Trajectory Repair

A.3.1 Entropy-Based Uncertainty Localization

We segment a failed trajectory $r^- = (s_1, \ldots, s_K)$ into semantic steps and compute token-level predictive entropy under the current student:

$$H_j = -\sum_{v \in \mathcal{V}} p_{\theta_t}(v \mid x, r_{<j}^-) \log p_{\theta_t}(v \mid x, r_{<j}^-). \tag{24}$$

Here $\mathcal{V}$ is the token vocabulary, and natural logarithms give entropy in nats. Step uncertainty combines the mean and upper quantile of token entropy, together with a local surge:

$$E_k^* = \lambda \ \text{Mean}_{j \in s_k} H_j + (1-\lambda) \ \text{P90}_{j \in s_k} H_j, \tag{25}$$

$$\Delta E_k = E_k^* - \frac{1}{m_k} \sum_{\ell=k-m_k}^{k-1} E_\ell^*, \qquad m_k = \min(m, k-1). \tag{26}$$

We set $\Delta E_1 = 0$. The weight satisfies $0 \le \lambda \le 1$; P90 is the within-step 90th percentile. Using Min-Max normalization with $\lambda = 0.5$, $m = 3$, and $w_\ell = w_r = 1$ (constant profiles map to zero), the strongest combined signal determines the candidate anchor and its surrounding window:

$$k^* = \arg\max_k (\widetilde{E}_k^* + \widetilde{\Delta E}_k), \tag{27}$$

$$\mathcal{W}(k^*) = \{s_{\max(1, k^*-w_\ell)}, \ldots, s_{\min(K, k^*+w_r)}\}. \tag{28}$$

Entropy proposes candidate regions for auditing. The auditor may retain a valid high-entropy step or move the boundary earlier when a later uncertainty surge follows an unsupported premise.

A.3.2 Error Auditing and Type-Guided Hints

We use GPT-5.6-sol as the teacher model for error auditing and minimal-hint generation. The auditor combines the original input, failed trajectory, entropy profile, and medical tags to identify the error boundary and construct a recovery hint:

$$(e, b, h) = \mathcal{A}(x, r^-, \mathcal{E}(r^-), z^{\text{obs}}). \tag{29}$$

The auditor identifies the earliest consequential error e, selects a safe prefix boundary b, and constructs a minimal hint h conditioned on medical tags and the current error type. The hint translates the diagnosis into a concrete re-examination or reasoning operation, introduced after the safe prefix by a natural transition followed by an error-conditioned cue, for the same student to continue. The auditor receives no reference answer, and the hint supplies neither the final diagnosis nor an answer option. Error labels guide hint construction but are not shown directly to the student.

A.3.3 Multiple Same-Model Continuations

We generate eight rollouts per question and allow at most five hint-guided continuation attempts for a failed trajectory. Trajectories generated and repaired by M_t form training data for the next round, M_{t+1}. The trajectory is decomposed into a retained prefix and discarded suffix,

$$r^- = p_b \oplus s^-, \qquad p_b = (s_1, \ldots, s_b). \tag{30}$$

The prefix is retained verbatim, and the same student that generated the failure continues from the prefix and recovery cue:

$$s^+ \sim \pi_{\theta_t}(\cdot \mid x, p_b, g(h)). \tag{31}$$

The original question occupies the user message; $p_b \oplus g(h)$ prefills one unfinished assistant message. The structured audit remains metadata outside the continuation input. Bounded retries regenerate the suffix while preserving the prefix. This restricts teacher intervention to diagnosis and guidance and leaves the continuation to the student's own policy.

B Supplementary Experiments and Data

B.1 Learning Value of Repair Supervision and Evolution Direction

MedEvo depends on selecting valuable trajectories, providing effective correction signals, and identifying what to learn next, beyond repair success alone. We examine these three questions below.

Which trajectories are worth repairing? Long failed trajectories may retain useful intermediate reasoning. We compare the longest and randomly selected 8,000 questions in single-round continuation training. The longest group has lower repair success (55.70% vs. 67.50%) but higher downstream accuracy (66.78% vs. 64.35%), motivating selection by learning value as well as repairability. See Figure 5(a).

How do hints provide learning signals? Hints direct the student to reconsider evidence. We compare 1, 3, 5, and 8 hint-guided continuation attempts to test whether more correction opportunities improve supervision. Mean accuracy rises from 65.78% to 67.01%, then falls to 66.74% and 64.10%. This non-monotonic pattern motivates bounded attempts with quality screening. See Figure 5(b).

Which feedback guides effective evolution? We compare training-error statistics with a fixed, independent 5,000-question feedback set as sources for prioritizing next-round repairs.

As shown in Figure 5(c,d), training-error feedback raises training accuracy from 63.23% to 78.00% but reduces benchmark accuracy from 65.57% to 64.01%; fixed-set feedback reaches 71.00% and 69.12%, respectively. Training trajectories provide repair candidates, while independent feedback diagnoses weaknesses beyond those questions and guides candidate priority.

Figure 5 summarizes these three comparisons.

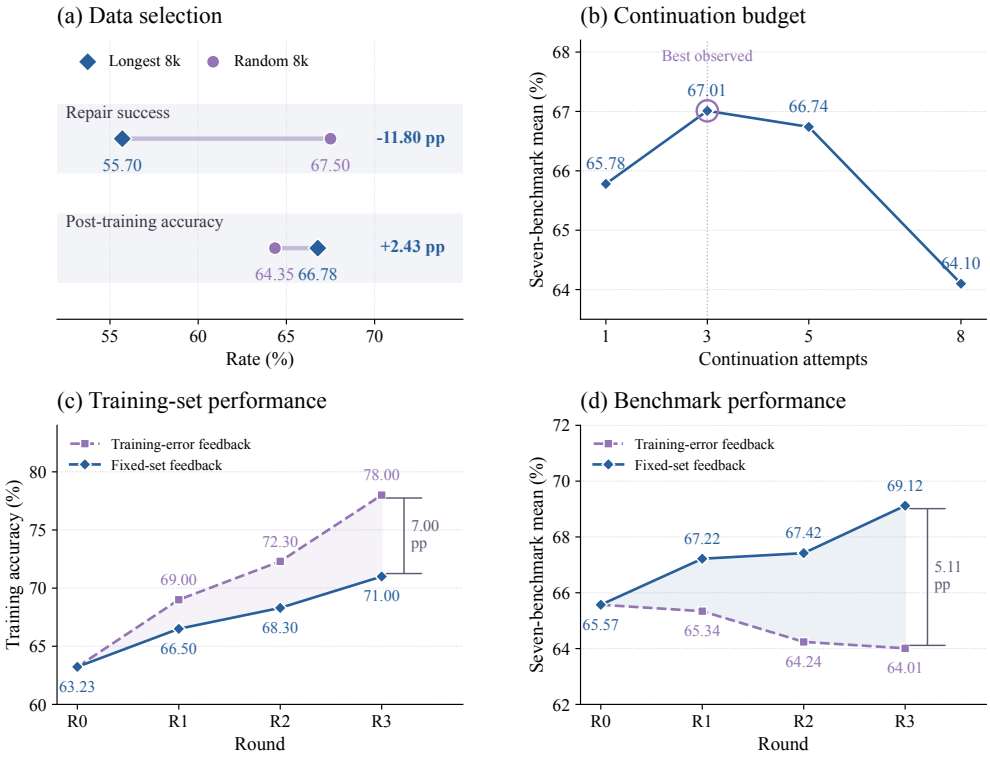


Figure 5: Data selection, continuation budget, and feedback direction. (a) Repair success and downstream means for longest-8k and random-8k selection. (b) Post-training means under different hint-guided continuation budgets. (c,d) Training accuracy and independent seven-benchmark means under two feedback sources; held-out feedback examples are excluded from training.

Additional repair analysis. Figure 6 links model-scale gains to repair behavior. With entropy localization, acceptance increases from 25.2% to 40.9% as the continuation budget grows from one to eight attempts, compared with 19.8% to 35.8% without localization. The erroneous-step rate rises from 16.2% to 64.1% across increasing entropy bins, supporting uncertainty-guided inspection. Across difficulty levels L1–L5, MedEvo improves accuracy over Base, with gains of 3.8, 4.9, 5.6, 7.2, and 8.7 percentage points. Acceptance measures the proportion of retained repair candidates.

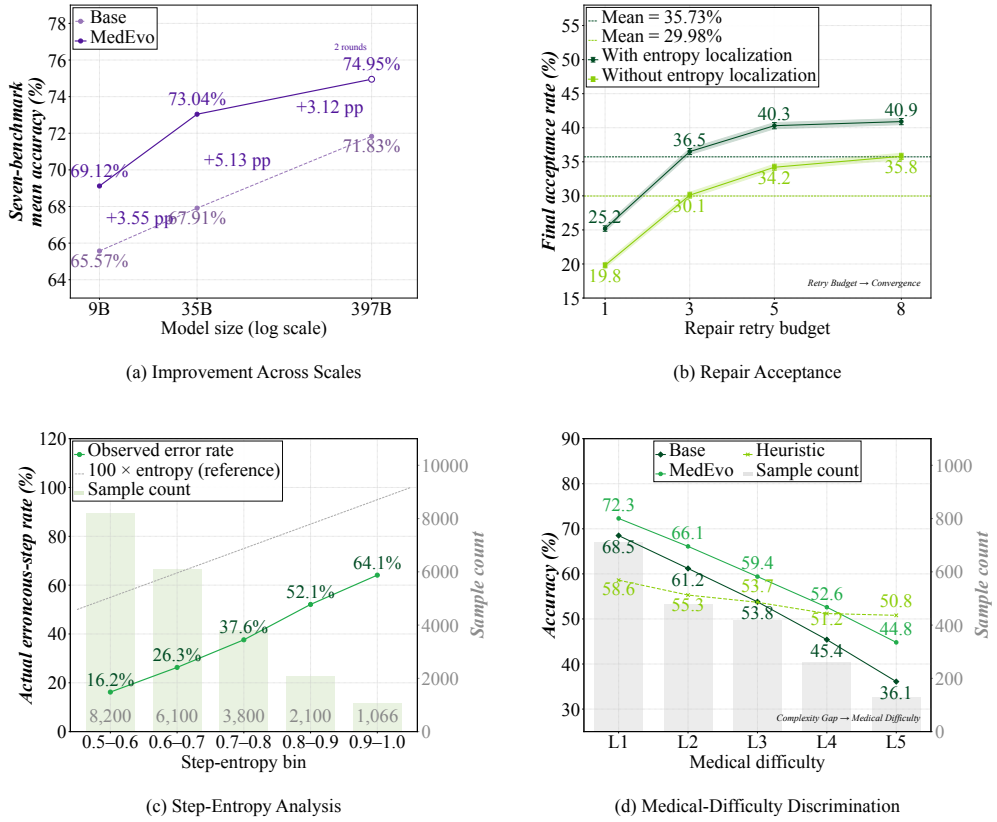
B.2 Additional Experimental Details

B.2.1 Paired Bootstrap Confidence Intervals

Table 7: Paired bootstrap confidence intervals for 9B gains across seven benchmarks.

Benchmark	Base (%)	MedEvo (%)	Δ (pp)	95% CI	Significant
PMC-VQA	59.20	62.68	+3.48	[+1.9, +5.1]	Yes
VQA-RAD	68.10	70.40	+2.30	[+0.8, +3.8]	Yes
SLAKE	76.80	78.23	+1.43	[+0.2, +2.7]	Yes
PathVQA	48.80	53.34	+4.54	[+2.6, +6.5]	Yes
OmniMedVQA	82.20	85.41	+3.21	[+1.5, +4.9]	Yes
MedXpertQA-MM	47.40	54.50	+7.10	[+5.2, +9.0]	Yes
MMMU-Med	76.50	79.28	+2.78	[+1.3, +4.3]	Yes
Macro average	65.57	69.12	+3.55	[+2.4, +4.7]	Yes

Paired bootstrap uses 10,000 question-level resamples. The macro-average is obtained by independently resampling within each benchmark and averaging the seven accuracies. Gains are significant when their 95% confidence intervals exclude zero.



(a) Improvement Across Scales

(b) Repair Acceptance

(c) Step-Entropy Analysis

(d) Medical-Difficulty Discrimination

Figure 6: Analysis across scale, repair budget, entropy, and difficulty. (a) Seven-benchmark means using 9B/35B R3 and 397B R2. (b) Repair acceptance with and without entropy localization; intervals are pointwise 95% Wilson intervals using 45,000 attempts per setting. (c) Erroneous-step rates and step counts by entropy bin; the diagonal provides a numerical reference. (d) Accuracy by medical difficulty, with sample counts on the right axis.

B.2.2 Human Audit and Reliability Analysis

Table 8 reports the five audit dimensions and inter-rater agreement. Hints pass the no-leakage check

Table 8: Human audit results. V, R, D, and K denote visual, reasoning, differential-diagnosis, and knowledge errors; pass rates are percentages.

Audit dimension	Overall	κ (95% CI)	V	R	D	K
Original trajectory is erroneous	95.5	0.90 (0.83–0.97)	96	96	94	96
Valid repair boundary	84.5	0.85 (0.66–0.88)	88	84	82	84
No answer leakage in hint	97.2	0.94 (0.87–1.00)	96	100	96	97
Repaired reasoning grounded in images	78.5	0.70 (0.58–0.82)	76	80	76	82
Correct repaired answer	88.3	0.89 (0.81–0.97)	88	89	88	88

in 97.2% of cases, and repair boundaries pass in 84.5%. Correct repaired answers reach 88.3%, while visual grounding passes in 78.5%.

B.2.3 Stability across Random Seeds

Table 9 reports seven-benchmark mean accuracy for three independent 9B training runs, using sample standard deviation. Base is the same fixed initial model. R3 reaches $69.09 \pm 0.24\%$, a mean gain of 3.52 points over Base.

Table 9: Results across random seeds (%).

Model	Seed 1	Seed 2	Seed 3	Mean ± SD	Source
Base	65.57	65.57	65.57	65.57 ± 0.00	Fixed initialization
MedEvo R1	67.22	67.05	67.38	67.22 ± 0.17	3 independent runs
MedEvo R2	67.42	67.21	67.60	67.41 ± 0.20	3 independent runs
MedEvo R3	69.12	68.84	69.31	69.09 ± 0.24	3 independent runs

Table 10: Question-level construction counts. When repair fails, an existing correct rollout for that question is used; final training-set size remains fixed.

Generation → Training	Question bank	Attempted	Repaired	Admitted	Mean attempts
$R0 \rightarrow R1$	81293	27233	19567	81293	3.37
$R1 \rightarrow R2$	81293	25831	18258	81293	2.48
$R2 \rightarrow R3$	81293	23620	16797	81293	2.12

B.2.4 Round-wise Data Construction

Each repair allows at most five continuation attempts. Mean attempts decrease from 3.37 to 2.48 and 2.12 over R0→R1, R1→R2, and R2→R3, indicating improved hint-guided correction efficiency across rounds.

Each training set contains 81,293 questions. Attempted repairs decline from 27,233 to 23,620, and successful repairs from 19,567 to 16,797.

B.2.5 Experimental Setup

Training and data. We use supervised fine-tuning from a shared Base, with an 81,293-question training bank and a separate 5,000-question feedback set. Correct rollouts and accepted, hint-guided suffix repairs form the supervision; repair filtering uses D_{repair} and Δ_{repair}. Data splits and training settings appear in Appendix C.1.5.

Compared methods. Baselines include HHD(Wang et al., 2026b) (Baseline 1), TrajFusion(Deng et al., 2026) (Baseline 2), and teacher-distillation SFT using supervision generated by Qwen-3.8-Max (Baseline 3). The 9B ablations compare repair designs and feedback configurations; 35B/397B assess scale, with additional selection and continuation-budget experiments.

Evaluation. MenUniEval evaluates PMC-VQA, VQA-RAD, SLAKE, PathVQA, OmniMedVQA, MedXpertQA-MM, and MMMU-Med using a shared answering prompt and parser. We report benchmark accuracy and its seven-task arithmetic mean, with repair success reported separately.

B.2.6 Evolution across Twenty Error Subtypes

Figure 7 expands the four families into twenty subtypes. Bar heights are normalized by each subtype's R0 rate, while in-bar labels retain the reported percentages. All subtypes are below R0 at R3. The knowledge-gap family has an R3/R0 ratio of approximately 0.89, compared with 0.72–0.75 for the other families, indicating a smaller relative reduction in this breakdown. Rates are reported separately by error subtype.

B.2.7 Round-wise Feedback Statistics

B.3 Detailed Error Statistics for the Current Rounds

Percentages are retained as supplied.

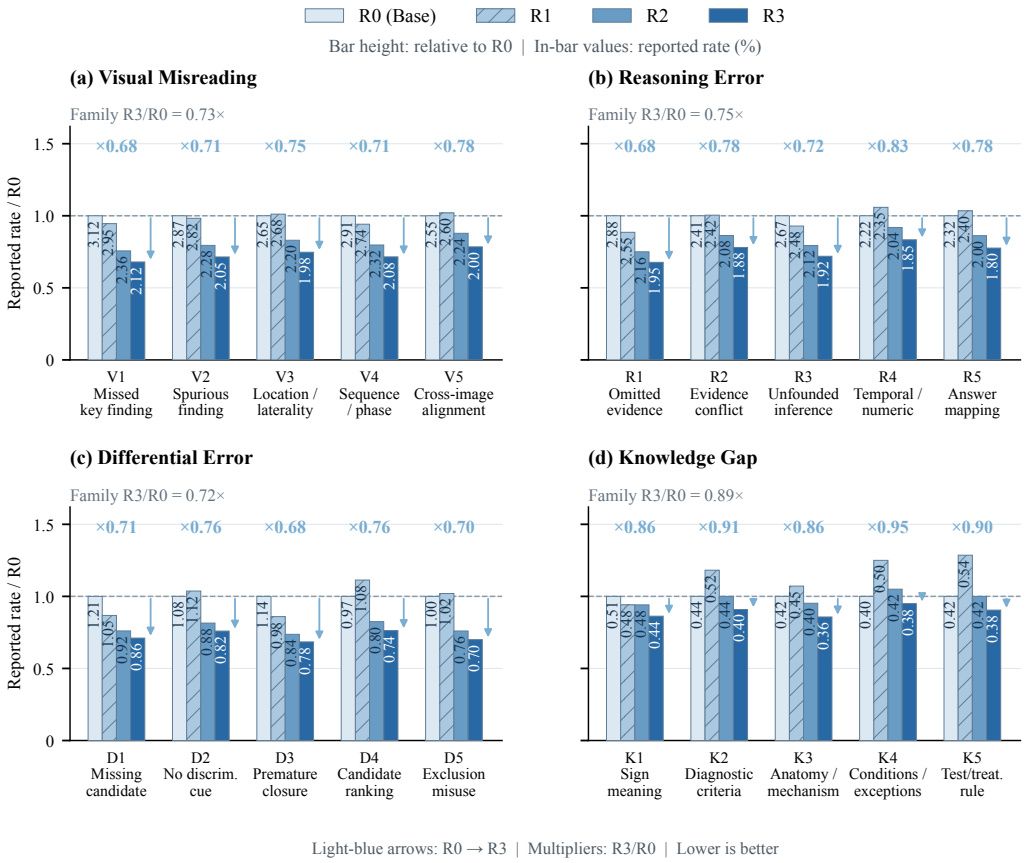


Figure 7: Twenty error subtypes of the 9B model across R0–R3. Each panel contains five subtypes. Bar heights divide reported rates by the corresponding subtype's R0 rate; in-bar labels give original percentages. Light-blue arrows and multipliers show R3 relative to R0, with lower values indicating larger relative reductions. Family R3/R0 uses sums of subtype rates, not an average of subtype ratios. Horizontal codes such as R1–R5 denote reasoning-error subtypes, distinct from round labels R0–R3 in the legend.

Table 11: Modality error rates (%) on the fixed feedback set, with group sizes.

Modality	N	R0	R1	R2	R3
Radiography	1509	37.18	35.98	31.28	32.21
Nuclear medicine	234	34.62	26.50	24.79	22.65
Ultrasound	351	44.73	39.03	39.60	36.18
Endoscopy	247	32.39	30.36	27.13	21.86
Pathology	378	34.13	34.66	30.95	28.31
Ophthalmology	135	34.07	27.41	22.96	20.00
Dermatology	593	30.35	27.82	24.62	23.10
Dental	308	40.26	37.01	37.66	32.79
Functional tests	1245	20.00	17.75	16.22	15.50

Table 12: Round-wise error-family percentages, summed over the five subtypes in each family. Subtype details are in the appendix. This breakdown is reported separately from modality error rates.

Error family	R0	R1	R2	R3
Visual misreading	14.10	13.79	11.40	10.23
Reasoning error	12.50	12.20	10.40	9.40
Differential error	5.40	5.25	4.20	3.90
Knowledge gap	2.19	2.49	2.16	1.96

Table 13: Percentages for the 20 error subtypes.

Subtype	R0	R1	R2	R3
V1 Missed key finding	3.12	2.95	2.36	2.12
V2 Spurious finding	2.87	2.82	2.28	2.05
V3 Location / laterality	2.65	2.68	2.20	1.98
V4 Sequence / phase	2.91	2.74	2.32	2.08
V5 Cross-image alignment	2.55	2.60	2.24	2.00
R1 Omitted evidence	2.88	2.55	2.16	1.95
R2 Conflicting evidence	2.41	2.42	2.08	1.88
R3 Unsupported inference	2.67	2.48	2.12	1.92
R4 Temporal / numeric	2.22	2.35	2.04	1.85
R5 Answer mapping	2.32	2.40	2.00	1.80
D1 Missing candidate	1.21	1.05	0.92	0.86
D2 No discriminating cue	1.08	1.12	0.88	0.82
D3 Premature closure	1.14	0.98	0.84	0.78
D4 Candidate ranking	0.97	1.08	0.80	0.74
D5 Exclusion misuse	1.00	1.02	0.76	0.70
K1 Sign interpretation	0.51	0.48	0.48	0.44
K2 Diagnostic criteria	0.44	0.52	0.44	0.40
K3 Anatomy / mechanism	0.42	0.45	0.40	0.36
K4 Conditions / exceptions	0.40	0.50	0.42	0.38
K5 Test / treatment rule	0.42	0.54	0.42	0.38

Table 14: Attempted and successful repairs by error family.

Round / family	Attempted	Successful	Success (%)
R1 / Visual misreading	12189	9826	80.61
R1 / Reasoning error	9355	5279	56.43
R1 / Differential error	4064	2995	73.70
R1 / Knowledge gap	1625	1467	90.28
R2 / Visual misreading	10245	6789	66.27
R2 / Reasoning error	7857	6246	79.50
R2 / Differential error	6742	4673	69.31
R2 / Knowledge gap	987	550	55.72
R3 / Visual misreading	10478	7094	67.70
R3 / Reasoning error	6789	5647	83.18
R3 / Differential error	4786	3291	68.76
R3 / Knowledge gap	1567	765	48.82

C Medical Taxonomy and Error Categories

C.1 Taxonomy, Permissions, and Execution Contracts

C.1.1 Four-Dimensional Medical Tag Dictionary

The medical taxonomy describes each question along four independent axes: modality, anatomy, disease, and difficulty. Hierarchical labels support coarse-to-fine capability analysis and training-question retrieval; a question may carry multiple labels.

Modality hierarchy.

Parent	Second-level scope
1 Radiography	X-ray; CT; MRI; mammography; fluoroscopy; angiography
2 Nuclear medicine	PET-CT; SPECT; radionuclide scans
3 Ultrasound	B-mode; echocardiography; vascular; interventional
4 Endoscopy	Gastrointestinal; bronchial; laryngeal/nasal; bladder; joint; laparoscopic
5 Pathology	Histology; cytology; molecular pathology
6 Ophthalmology	Fundus photography; OCT; FFA/ICGA; slit lamp
7 Dermatology	Dermoscopy; clinical photography; reflectance confocal microscopy
8 Dental	Panoramic; CBCT; periapical; cephalometric
9 Functional tests	ECG; EEG; EMG

Anatomy hierarchy.

Parent	Second-level scope
1 Nervous system	Brain; spinal cord; peripheral nerves
2 Head and neck	Orbit; ear; sinuses; oral/maxillofacial; pharynx/larynx; neck soft tissue
3 Thorax	Lung/airway; mediastinum; pleura; heart; great vessels; breast; chest wall
4 Abdomen	Hepatobiliary; pancreas; spleen; gastrointestinal; renal/urinary; retroperitoneum
5 Pelvis	Male reproductive; female reproductive; pelvic floor; pelvic bones
6 Spine	Cervical; thoracic; lumbar; sacrococcygeal; disc; canal; paraspinal tissue
7 Musculoskeletal	Craniofacial bones; upper limb; lower limb; muscles; periarticular soft tissue
8 Vascular	Arteries; veins; lymphatic system
9 Skin and soft tissue	Head/face; trunk; limbs; subcutaneous; deep soft tissue

Disease hierarchy.

Parent	Second-level scope
1 Neoplasm	Malignant; benign; metastatic; precancerous categories
2 Infection and inflammation	Bacterial; tuberculosis; viral; fungal; parasitic; noninfectious inflammation
3 Vascular disease	Ischemic; hemorrhagic; aneurysm; thrombosis/embolism; malformations; atherosclerosis/stenosis
4 Degenerative disease	Joint; neural; organ degeneration
5 Trauma	Fracture; joint injury; visceral injury; head injury; foreign body
6 Congenital/developmental	Cardiovascular; nervous; skeletal; digestive; urinary; respiratory
7 Metabolic/endocrine	Diabetes complications; thyroid; adrenal; gout; metabolic bone
8 Autoimmune	Lupus; ankylosing spondylitis; multiple sclerosis; NMOSD; autoimmune liver; IgG4; vasculitis; systemic sclerosis; myositis; Sjogren
9 Functional disorders	Obstruction; effusion; ectopic gas; torsion; stones
10 Iatrogenic/postoperative	Normal postoperative change; complications; implants; post-radiotherapy; drug-related

Leaf specificity and multi-label assignment. CT branches into noncontrast, contrast-enhanced, CTA, HRCT, and low-dose CT; MRI includes T1, T2, FLAIR, DWI, MRA, fMRI, MRS, and SWI. Anatomy can descend from thorax to lung/airway and then to a lung lobe, or from abdomen to pancreas and then head/body/tail. Disease labels descend from a broad family to a mechanism or disease group and then to a named entry. Preserve image IDs for mixed modalities and record every supported anatomical site. Assign the deepest supported node, not a guessed leaf; use unknown, not-applicable, or a supported parent when necessary. Expand ancestors as a set and count each question once per group. Related concepts at different paths require an explicit alias ledger, not automatic merging. Store source text, evidence span/image ID, confidence, visibility, and taxonomy version for each assignment.

Difficulty rubric.

Level	Task requirement
L1	Single modality and conspicuous canonical findings; direct recognition.
L2	Common presentation requiring localization or characterization and limited context.
L3	Atypical or multiple findings requiring sequence integration or differential comparison.
L4	Subtle findings, multisystem involvement, or complex context requiring extensive integration.
L5	Highly complex or disputed tasks requiring multidisciplinary, quantitative, or specialized knowledge.

Assign difficulty from the actual question: visibility, number of required comparisons, evidence integration, and specialized knowledge. A disease name, modality, or a single failed rollout does not determine the level. Rare disease alone does not force L5. Record the rationale and adjudicate ambiguous adjacent levels.

C.1.2 Error Codebook: Four Parents and Twenty Subtypes

We specify five subtypes under each of the four existing parent categories. Each row defines an observable diagnostic criterion and a navigation operation; it does not supply a gold answer.

V: Visual misreading.

Code / subtype	Diagnostic criterion	Navigation operation
V1 Missed finding	A supported salient finding is not observed or described.	Reinspect the relevant region.
V2 Hallucinated or misidentified finding	An artifact, normal structure, or unsupported feature is read as a lesion.	Check direct visual support.
V3 Anatomy or laterality error	The organ, level, compartment, or side conflicts with visible evidence.	Recheck landmarks and orientation.
V4 Modality, sequence, or phase error	Acquisition type, sequence, or phase is misidentified, distorting feature reading.	Verify acquisition context before reading features.
V5 Cross-image correspondence error	A lesion is matched incorrectly across images, slices, or views.	Align images and anatomical locations.

R: Reasoning error.

Code / subtype	Diagnostic criterion	Navigation operation
R1 Omitted evidence	An observed finding or stated fact is omitted from inference.	Reintroduce the omitted evidence.
R2 Conflicting evidence integration	Correctly extracted evidence conflicts and is not reconciled.	Reconcile supporting and opposing evidence.
R3 Unsupported inference or causal leap	A conclusion lacks the required logical or causal link.	Restore a testable intermediate inference.
R4 Temporal or quantitative error	Correctly read facts are combined with an incorrect temporal or numerical relation.	Check chronology, units, or calculation.
R5 Task constraint or answer-mapping error	The reasoning violates the requested target, negation, or option mapping.	Recheck the requested target and answer mapping.

D: Wrong differential.

Code / subtype	Diagnostic criterion	Navigation operation
D1 Omitted candidate	A necessary evidence-compatible alternative is not considered.	Broaden the candidate comparison.
D2 Missing discriminative comparison	Alternatives are listed without comparing discriminative evidence.	Identify discriminative evidence.
D3 Premature closure or anchoring	An initial diagnosis is fixed before counterevidence is considered.	Revisit counterevidence and alternatives.
D4 Incorrect candidate ranking	Known support, counterevidence, or priors are weighted incorrectly.	Compare the strength of evidence.
D5 Misapplied exclusion	An alternative is excluded using insufficient negative evidence or a nonessential criterion.	Check whether exclusion is justified.

K: Knowledge gap.

Code / subtype	Diagnostic criterion	Navigation operation
K1 Missing finding interpretation	The required medical meaning of a finding or term is missing.	Cue the relevant conceptual relation.
K2 Missing diagnostic criterion	A required diagnostic, grading, or staging rule is unavailable.	Cue the relevant rule category.
K3 Missing anatomy or pathophysiology	Knowledge linking anatomy, mechanism, and manifestations is missing.	Bridge the minimal mechanistic relation.
K4 Missing applicability condition or exception	A general rule is known but its prerequisites or exceptions are missing.	Cue the applicability check.
K5 Missing investigation or management knowledge	A next-investigation or management task requires an unavailable principle.	Cue the decision principle without the answer.

Boundary rules and reporting. A finding not perceived is V1; a finding already perceived but omitted from inference is R1. Incorrect correspondence across images is V5; conflicting correctly extracted evidence is R2. Candidate-set construction, comparison, and exclusion belong to D; a demonstrated missing prerequisite rule belongs to K. Knowledge-gap labels require evidence of a missing prerequisite rule. Select the earliest consequential supported error as the primary subtype and record later failures as secondary labels, with step spans and evidence. If the parent is clear but the subtype is not, retain the parent and an unresolved-subtype status; if even the parent is unsupported,

mark the audit unresolved. These are annotation states, not extra error classes. Use one primary subtype per trajectory for additive counts; report secondary labels separately. Aggregate V1–V5 into V, and similarly for R, D, and K. Re-audit fresh rollouts at each checkpoint. Subtype rates use subtype-specific annotations and denominators.

C.1.3 Information Access Matrix

Component	Permitted inputs	Excluded inputs
Feedback evaluator	Feedback items, labels, current predictions; emits aggregates and transitions	Final-test results for planning; feedback export into SFT
Planner / retrieval	Aggregate feedback, training tag metadata (including true disease labels), coverage, budgets	Final-test items/results; any authority to forward answer-derived labels to repair
Auditor / hint renderer	Original images, question/options, current failed trajectory, entropy profile, observable tags with provenance	Gold answer, answer-derived disease/location, reference rationale, hidden retrieval labels
Continuation M_t	Original input and assistant prefill $p_b \oplus g(h)$	Gold answer, full audit, hidden tags, teacher solution
Admission / exporter	Candidate, reference answer for verification, boundary, provenance, suffix scores	Feedback/test examples as training targets
SFT trainer	Frozen training manifest; $p_b \oplus s^+$, masks	Construction hint and full audit as model inputs; feedback/test samples

Use explicit allowlist serialization, not deletion of one answer field from an otherwise unrestricted object. Reject fields without observable provenance, including filenames/IDs encoding labels. An authorized evaluator can identify a wrong training answer before repair without transmitting the answer. Generation sees only a boolean routing decision, not the key. Record hashes of redacted requests for leakage audit. The planner's chosen hint family is passed through an eligibility check derived from observable context and current diagnosis; disease-bearing planner explanations are never forwarded.

C.1.4 Structured Fields and Plan Validation

The execution interface requires JSON records containing the following fields. An omitted measurement is null with a reason; zero means measured zero.

Question metadata: canonical question ID; source ID; patient/case/article cluster IDs when available; image hashes and near-duplicate cluster; split; taxonomy version; per-image modality; all four tag sets; per-tag provenance, visibility (statistical or observable), and uncertainty.

Feedback record: checkpoint hash; split hash; group and resolved parent; unique sample count; error count/rate; all four transition counts; both regression denominators; current error-type/subtype counts; confidence interval; training-pool unique availability; measured repair/admission/cost fields.

Plan: round ID; generator and initialization hashes; planner type/version and seed; feedback hash; target group IDs; numeric feedback evidence and fallback ancestry; training sources; selected IDs; attempt/call/token ceilings; per-group quota; permissible hint operation; history replay allocation; replacement rule; supervised-token cap; shortfall/reallocation policy; fixed training-config hash; acceptance-validator result.

Repair record: question/checkpoint IDs; fresh rollout hash; semantic steps and token spans; entropy profile/proposal; current error type/subtype; audit boundary and exact prefix span; observable-tag payload hash; first-person cue; each continuation and seed; correctness and integrity checks; current-model suffix scores and cutoffs; admission/rejection reason; all generated/scored tokens.

Mix and training record: ordered example IDs; source round and generator; dedup/replacement ledger; historical/new role; token masks; sequence lengths; unit weights; packing map; effective repetitions and exposures; optimizer and scheduler reset flags; learning rate/warmup; initialization hash; output hash; second-stage and cumulative costs.

Execution receipt: planned versus realized quotas, unique selected and attempted failures, successful and admitted repairs, substitutions, exhausted budgets, rejected records, uncovered groups, and actual training tokens.

The validator rejects cross-split IDs, duplicate quota charges, budgets above caps, out-of-pool retrieval, unsupported hint operations, missing visibility provenance, or changed training hyperparameters. Save the validated plan before generation and record actual execution counts afterward.

The complete feedback-agent prompt is given in Appendix D.2.

C.1.5 Training Configuration and Evaluation Scope

Table 15: Training configuration

Configuration	Value
Framework	ms-swift + Megatron-LM backend
Precision	bf16
micro_batch_size	1
global_batch_size	64
packing	true
attention_backend	flash
calculate_per_token_loss	true
freeze_llm / freeze_vit / freeze_aligner	false / true / true
optimizer_cpu_offload	true
use_precision_aware_optimizer	true
offload_fraction	1.0
moe_grouped_gemm	true
moe_shared_expert_overlap	true
moe_aux_loss_coeff	1e-6
lr	1e-5
lr_warmup_fraction	0.05
min_lr	1e-7
add_non_thinking_prefix	true
loss_scale	ignore_empty_think
MAX_PIXELS	3145728
dataset_num_proc	96
dataloader_num_workers	8
logging_steps	1
eval_iters	1000
recompute_granularity	selective
sequence_parallel	true
max_length	28672
epochs	2
save_steps	1000 (pilot: 100)
save_total_limit	2
Recompute modules	core_attn + moe + mlp
moe_permute_fusion	true
Parallelism	TP4/EP8/PP1/CP1/ETP1 $\rightarrow$ DP2 on 64 GPUs

Repaired examples use suffix-only loss, while original correct examples use full-response loss; navigation hints are removed before SFT.

Section 4 reports seven-benchmark accuracy and its arithmetic mean, comparing each backbone's best complete checkpoint with its corresponding Base.

D Prompts

D.1 Shared Question-Answering Prompt

All question-answering tasks use the following instruction verbatim.

Shared Question-Answering Instruction

Please reason step by step, and put your final answer within `\boxed{}`.

D.2 Main-Agent Feedback and Next-Round Planning

The following template connects feedback diagnosis to next-round selection and repair priorities under the method protocol. It specifies agent behavior; run-specific inputs and execution records are required to document its use.

Main Agent: Feedback Summary and Next-Round Data Plan

You are the feedback and data-planning agent for MedEvo. Diagnose the current model's weaknesses and produce an executable next-round data plan. Do not write solutions to training questions or change the experimental comparison settings.

Inputs. Current round and checkpoint; the fixed 5,000-question feedback-set results for current and previous checkpoints; four-axis medical tags; four error families and their twenty subtypes; per-group sample counts, correctness transitions, and current error diagnoses; available training-question IDs and tags; historical trajectory records; fixed generation/repair budgets, support threshold, regression weight α, and admission rules. Missing inputs must be reported, never invented.

1. Summarize feedback. For each supported tag group, report error rate, change from the previous round, corrected errors, regressions, and persistent errors. Summarize four-family and twenty-subtype distributions with explicit denominators. Count each question once within a group; do not sum overlapping groups into an overall rate. Merge unsupported groups into supported parents. Distinguish observed changes from explanations that require further evidence.

2. Set priorities. Rank groups using $S_{g,t} = \epsilon_{g,t} + \alpha N^{10}_{g,t}/n_g$ under the supplied configuration. Identify dominant or persistent error types within each priority group. State which measured feedback supports each priority; do not infer that a high-error modality contains every error subtype.

3. Select training questions. Allocate the fixed budget proportionally to group scores, cap quotas by available unique questions, and redistribute unused quota. Preserve modality coverage and apply the configured fallback when all scores are zero or support is insufficient. Retrieve by medical tags, using historical errors only as prioritization cues. Deduplicate multi-tag questions and return selected training IDs, group quotas, and any shortfall. Never select feedback or final-test examples.

4. Guide fresh rollouts and repair. Ask the current checkpoint to generate the configured eight rollouts per selected question. Diagnose their actual errors again. Use feedback-identified modality–error pairs to prioritize failed training trajectories: prefer joint matches, followed by modality-only or error-type-only matches when necessary. For each selected trajectory, pass its observable medical tags and current error diagnosis to the hint generator. The recovery hint must address the error actually observed in this trajectory. Send observable tags and local diagnoses to the auditor and hint generator, without reference answers or answer-bearing metadata. Retain correct rollouts and accepted continuations under the existing screening rules; if repair fails, use an available quality-screened correct rollout. If none is available, retain the valid historical target or report no accepted target.

5. Plan the next round. Keep at most one active target per question and replace it only with an accepted improvement. Train on the complete updated dataset from M_0. Keep training settings and evaluation protocol fixed. Evaluate the next checkpoint on the same feedback set to measure the targeted changes; independent benchmark results are reported separately and must not drive selection.

Output. Return a structured plan containing: round and checkpoint; feedback summary and denominators; priority tag/error pairs and supporting statistics; score and quota per group; selected unique training IDs; rollout and repair budgets; repair priorities; dataset-retention/replacement rules; initialization M_0; next-round feedback comparisons; and missing inputs or budget shortfalls. After execution, record actual generated, repaired, admitted, and replaced counts.

D.3 Stage 1 Audit Prompt

The diagnostic model receives the image, the question with options, the failed trajectory, the complete step-level uncertainty profile, and the entropy-derived anchor and candidate window. The auditor is gold-blind: the reference answer is not an input to diagnosis. Answer-derived disease labels, hidden lesion positions, reference explanations, and planner retrieval metadata are also excluded.

Only provenance-checked observable tags are allowed. The entropy proposal is a localization prior, not an error label: the auditor inspects the proposed region first, traces its evidence dependencies backward, and may move the final boundary when the medical or logical evidence requires it. It returns a structured audit that identifies a compact revision region and constrains the subsequent continuation. The emitted hint must remain answer-blind. Placeholders are shown in braces and are filled at runtime.

Prompt D.3: Stage-1 Entropy-Guided Medical Trajectory Audit

You are a senior medical reviewer auditing a failed multimodal reasoning trajectory segmented into semantic steps $s_1, \ldots, s_K$.

You are given:
(1) the medical image(s);
(2) the clinical question and answer options;
(3) the failed student trajectory segmented into semantic steps;
(4) STEP_ENTROPY: the uncertainty score E_k^* of every step;
(5) NORMALIZED_ENTROPY: the normalized uncertainty score of every step;
(6) LOCAL_SURGE: the local increase ΔE_k relative to the preceding step history;
(7) ENTROPY_ANCHOR: the step proposed by the uncertainty locator;
(8) ENTROPY_WINDOW: the compact candidate window surrounding that anchor, and ENTROPY_THRESHOLD if a threshold was used.
(9) OBSERVABLE_TAGS: per-image modality and other context supported by the original input, with evidence provenance; never hidden answer-derived tags.

The entropy signal guides audit localization. High uncertainty can reflect a genuine failure, but it can also arise from difficult terminology, ambiguous visual evidence, or a medically valid differential diagnosis. Combine the entropy proposal with the image, clinical evidence, and dependencies between reasoning steps to identify the first consequential failure.

Error-conditioned hint construction. Construct the hint from OBSERVABLE_TAGS, ERROR_TYPE, ERROR_SUBTYPE, and KEY_ERROR. Select a recovery operation addressing the diagnosed failure:

- **Visual misreading:** direct attention to the relevant region, visual attribute, or correspondence across available images.
- **Reasoning error:** revisit an unsupported inference or reconnect observed evidence to the requested decision.
- **Differential diagnosis error:** compare discriminative evidence or reconsider an insufficiently examined alternative.
- **Knowledge gap:** identify the relevant concept, rule, or applicability condition to reconsider.

Specify one concrete next operation grounded in the current question and trajectory. Express it as a brief recovery cue without supplying the corrected finding, final diagnosis, or answer option.
Audit procedure:
(1) Inspect ENTROPY_WINDOW first. Determine whether its local entropy peak or surge corresponds to an unsupported observation, invalid inference, incorrect differential comparison, or missing medical principle.
(2) Trace evidence dependencies backward from ENTROPY_ANCHOR. Decide whether the decisive failure begins inside the proposed window or an earlier step introduced the unsupported premise that caused the later entropy increase.
(3) Do not automatically select the highest-entropy step. Justify the final AUDIT_ANCHOR and SAFE_PREFIX_END using the medical image, clinical information, and logical dependencies.
(4) Prefer a compact localized repair and preserve every preceding step that remains medically and logically valid.
(5) You may move the audit anchor outside ENTROPY_WINDOW when the proposal misses the first consequential failure, but briefly explain why the entropy signal was misleading.

Return a JSON object with the following fields, grounded strictly in the provided image and text:

(a) ACTION: REPAIR or SKIP.
(b) STEP_SUMMARY: a brief one-sentence semantic summary for each step.
(c) ENTROPY_ASSESSMENT, containing: proposed_anchor, proposed_window, proposal_status, and brief_reason. proposal_status must be one of {HIT, PARTIAL_HIT, MISS, NON_ERROR_UNCERTAINTY}. HIT means that the first consequential failure is at or immediately adjacent to the proposed anchor; PARTIAL_HIT means that the proposal covers the propagated error region but the unsupported premise begins earlier; MISS means that the first failure lies outside the proposed window; and NON_ERROR_UNCERTAINTY means that the high-entropy region is medically or logically valid.
(d) AUDIT_ANCHOR: the index of the first consequentially incorrect step.
(e) SAFE_PREFIX_END, containing step_index and the exact verbatim_ending_span of the last fully valid step before the repair region.
(f) AUDIT_WINDOW: a short contiguous sequence of steps that should be reconsidered.
(g) ERROR_TYPE and ERROR_SUBTYPE: checkpoint-specific diagnosis, plus KEY_ERROR: the localized error stated without revealing the correct answer.
(h) HINT_TYPE: one of {perceptual_reexamination, inference_reconstruction, differential_disambiguation, knowledge_bridging}.
(i) MINIMAL_HINT: a brief recovery cue with flexible wording, conditioned on observable medical tags and the diagnosed local error. Name the evidence or reasoning operation to revisit. Do not include the error-category label or audit explanation. Introduce the cue with a natural transition; no fixed phrase or first-person form is required.
(j) IMAGE_CHECK, containing applicable (true or false) and target (the specific modality, anatomy, laterality, finding, relation, or measurement to re-inspect).
(k) KNOWLEDGE_SCOPE: the narrow medical principle to recall, or null.
(l) DIFFERENTIAL_FOCUS: the discriminative axis to compare, or null.
(m) STOP_CONDITION: the evidence condition under which the continuation should commit to an answer and stop.

Constraints:
- Do not access hidden disease labels or hidden lesion coordinates.
- Do not state a gold answer, corrected diagnosis, or correct option letter.
- Do not provide a replacement chain of thought or teacher-authored solution.
- Do not introduce symptoms, history, tests, measurements, or visual findings absent from the input.
- Do not discard a valid step merely because its entropy is high, and do not preserve an invalid step merely because its entropy is low.
- SAFE_PREFIX_END must precede the first consequential failure.
- AUDIT_WINDOW should normally overlap ENTROPY_WINDOW, but may deviate when supported by explicit medical or logical evidence.
- If only final-answer formatting or option selection is wrong and no reasoning revision is justified, set ACTION to SKIP.

D.4 Stage 2 Continuation Prompt

The continuation input contains the original question and an unfinished assistant message with the retained prefix and an error-conditioned recovery cue. Stage 1 constructs the cue from observable medical tags, ERROR_TYPE, ERROR_SUBTYPE, and KEY_ERROR. Feedback determines repair priority; the selected trajectory's own diagnosis determines cue content. The same student continues from this prefilled message.

Prompt D.4: Stage-2 Assistant-Prefill Continuation

User message:
{original medical image(s), question, and answer options}

Assistant message (unfinished):
{safe_prefix}
{Natural transition} {MINIMAL_HINT from Stage 1}

MINIMAL_HINT is inserted verbatim after the retained prefix. The student receives only the operational cue; structured error labels and the audit remain metadata. Generation continues immediately after the cue within the same assistant message, without repeating or modifying the prefix.

The SFT export removes the construction cue. Repaired examples use the suffix-only loss in Equation 6; original correct examples use full-response loss. Historical repairs retain their stored boundaries and masks. No audit or external hint is supplied at final evaluation. Record exported examples and token masks with the training configuration.

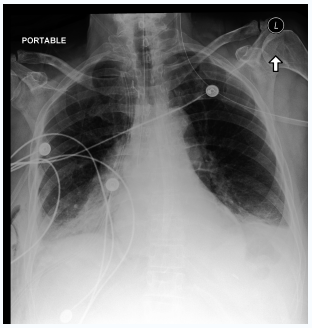
E Case Study

We present two medical image question-answering cases, including the question, options, and student model output, with the hint marked within the complete reasoning text.

E.1 Case 08: Chest radiograph

Original question and options

A 67-year-old man is recovering in the surgical intensive care unit following an esophagectomy. A portable anteroposterior chest radiograph (Figure A) is obtained to evaluate his postoperative status. The patient has a right-sided chest tube in place. Based on the radiographic findings, which of the following is the most likely cause of the bilateral lower lung opacities?
Options:
A. Pleural effusion
B. Pulmonary embolism
C. Emphysema
D. Cavitary tuberculosis
E. Tension pneumothorax
Please reason step by step, and put your final answer within \boxed{}.



Student model output

<think>

The user wants a step-by-step reasoning and a final boxed answer. Let me look at the radiograph carefully. It's a portable AP chest radiograph of a 67-year-old man post-esophagectomy. There's a right-sided chest tube. There are ECG leads overlying the chest. There's an NG tube or feeding tube coursing down the midline (post-esophagectomy, possibly gastric pull-up). The question asks about bilateral lower lung opacities. Looking at the image, the lower zones do not show convincing increased opacity; the costophrenic angles are not clearly blunted, and there is no definite meniscus-like contour. The hemidiaphragms remain relatively well defined. So pleural effusion is not strongly supported.

Hint

Before weighing the choices, re-check the morphology of the basal opacities: costophrenic-angle blunting, a meniscus-like contour, and partial obscuration of the hemidiaphragms would localize the process primarily to the...

Now, let me think about the other options: pulmonary embolism doesn't usually cause bilateral lower lobe opacities (may cause atelectasis or infarction, but not typical). Emphysema causes hyperlucency, not opacity. Cavitary TB causes upper lobe cavities, not bilateral lower lobe opacities. Tension pneumothorax causes hyperlucency and mediastinal shift, not opacity. So the answer is A, pleural effusion.

The question says "bilateral lower lung opacities" — in a postoperative patient with a chest tube, bilateral pleural effusions (or a chylothorax after esophagectomy) are common. The radiographic appearance of basal opacities with meniscus and blunting of the costophrenic angles supports pleural effusion.

Therefore the

Let me describe the radiographic findings in more detail: bilateral lower zone opacities with a meniscus sign, blunting of the costophrenic angles, obscuration of the hemidiaphragms; the chest tube is on the right; the left side shows a similar opacity. No evidence of hyperlucency or pneumothorax; no cavitary lesions; no findings of emphysema (which would be lucent). PE is not a radiographic diagnosis on chest X-ray anyway and wouldn't present as bilateral basal opacities.

Therefore the answer is A.

I should present step-by-step reasoning: identify the findings, exclude the options, and conclude with A.

Let me double-check the image description once more: the left costophrenic angle is blunted with a meniscus; the right base shows opacity consistent with effusion and possibly some atelectasis. The question asks for the "most likely cause" — pleural effusion.

Also, after esophagectomy, patients often develop pleural effusions due to chylothorax or transudate from fluid shifts. So A is clinically plausible as well.

Format: the answer should be moderately detailed, with step-by-step reasoning, and a final answer. Use headings or bullet points sparingly. Let me write a concise structured analysis: findings, differential exclusion, conclusion. Then give the final

Let me write it like this:

Step 1 — Systematic reading: portable AP film; cardiac silhouette is difficult to assess; bilateral lower zone opacities; blunting of the costophrenic angles with a meniscus-like upper border; obscuration of the hemidiaphragms; right-sided chest tube (likely for drainage of pleural fluid/air); no evidence of pneumothorax (no visceral pleural line with hyperlucency); no cavitary lesions; lung volumes are not hyperinflated.

Step 2 — Match with options:

- Pleural effusion: fits — dependent (basal) opacities with meniscus, blunting of the costophrenic angles, common after esophagectomy (including chylothorax).

- PE: not visible on CXR; wouldn't cause diffuse basal opacities.

- Emphysema: causes hyperlucency, flattened diaphragms — opposite of opacity.

- Cavitary TB: upper lobe cavitary lesions, not bilateral basal opacities.

- Tension pneumothorax: hyperlucent hemithorax, mediastinal shift away — not present.

Conclusion: A.

I might also mention that the chest tube on the right is for drainage of pleural fluid/air, and that the left effusion is likely also present.

Let me double-check the arrow in the image: it points to the left shoulder/clavicle area — probably just a marker. Not relevant.

Final

I'll write the response with clear reasoning. Format: short paragraphs or bullet points. Let me use a structured but concise approach</think>

Most appropriate answer: \boxed{A}.

E.2 Case 31: Elbow radiograph

Original question and options

A 32-year-old patient presents to the emergency department with acute right elbow pain, swelling, and limited range of motion following a fall onto an outstretched hand. An anteroposterior radiograph of the right elbow is obtained (Figure A). What is the most likely diagnosis?
Options:
A. Normal elbow radiograph
B. Isolated elbow dislocation
C. Elbow fracture
D. Advanced osteoarthritis
E. Osteochondritis dissecans
Please reason step by step, and put your final answer within \boxed{}.



Student model output

<think>

The user wants a step-by-step reasoning for a radiograph interpretation question. Let me carefully look at the image. The image is labeled ”R” (right) and shows an elbow radiograph. The question states it’s an AP radiograph, though the image looks somewhat like an oblique or lateral-ish view, but let’s focus on the pathology. Looking at the radiograph: I see the distal humerus, the proximal radius, and proximal ulna.

Hint

Before committing, trace the radial head and neck cortices to verify that the horizontal lucency causes true cortical interruption or fragment displacement rather than projectional overlap, then assess whether the radiocapitellar and ulnohumeral relationships are...

The radiocapitellar alignment appears normal (the radial head is opposite the capitellum), and the ulnohumeral articulation is also maintained, suggesting no dislocation. In case of an isolated dislocation, the radial head would be displaced from the capitellum. Here the alignment looks preserved, so an isolated dislocation is unlikely. The presence of a fracture line at the radial head suggests a radial head fracture, so the answer is C (elbow fracture). Osteoarthritis wouldn’t present acutely after trauma, and there are no osteophytes. Osteochondritis dissecans is a chronic condition with a loose body, not an acute trauma presentation.

Therefore, the most likely diagnosis is C: elbow fracture (radial head fracture). Let’s describe the reasoning steps:

1. Verify patient history: 32-year-old, fall on an outstretched hand (FOOSH) -> typical mechanism for radial head/neck fractures, coronoid fractures, lateral collateral ligament injuries, etc.

2. Radiographic findings: AP view of the right elbow shows a lucent line with cortical disruption at the radial head/neck, with some displacement/fragmentation - consistent with a radial head fracture. The radiocapitellar and ulnohumeral joints remain congruent, excluding dislocation. No evidence of advanced osteoarthritis (no joint space narrowing, osteophytes, subchondral sclerosis). No chronic loose body or cartilage defect suggestive of osteochondritis dissecans.

3. Therefore the answer is C.

Also, it’s worth mentioning that the ”fat pad sign” may be present (elevated anterior fat pad), indicating an intra-articular fracture/hemarthrosis, supporting the diagnosis of a fracture. In the image, there may be a faint anterior fat pad sign. But the key is the radial head fracture.

Let me double-check the image description once more: the image shows the humerus, ulna, and radius. In the proximal radius region, there’s a clear horizontal white line - a fracture of the radial head. The radial head may also be comminuted. So the answer is C.

Let me also consider the possibility of a coronoid fracture. Regardless, it’s an elbow fracture.

Thus, the final

Format: The answer should be moderately detailed, including step-by-step reasoning: mechanism, radiographic findings, exclusion of other options, conclusion. Use headings or bullet points. Final

Let me carefully describe the image: In the image, the radial head shows a transverse lucency consistent with a fracture; the radial head fragment is slightly displaced; the radiocapitellar relationship is preserved; the ulnohumeral angle is normal; no dislocation.

Let me also mention that the lateral fat pad sign supports an intra-articular fracture.

Let me write the response.

Also, the question states ”anteroposterior radiograph,” but the image is somewhat lateral; I won’t overemphasize the view.

Final

Step 1 – Mechanism of injury: A fall onto an outstretched hand (FOOSH) is the classic mechanism for a radial head/neck fracture, coronoid fracture, or lateral collateral ligament injury in a young adult.

Step 2 – Systematic review of the radiograph:

- **Joint alignment:** The radiocapitellar and ulnohumeral relationships are preserved – the radial head remains opposite the capitellum and the ulna articulates normally with the trochlea. This argues **against an isolated elbow dislocation** (option B), which would show obvious displacement of these relationships.

- **Bony detail:** There is a transverse lucent line with cortical disruption and a small displaced fragment at the radial head/neck region – i.e., a **fracture of the radial head**. A faint anterior fat□pad sign (sail sign) is also suggestive of an intra□articular fracture/hemarthrosis.

- **No degenerative changes:** There is no joint space narrowing, osteophyte formation, or subchondral sclerosis, excluding **advanced osteoarthritis** (option D), which would also be unlikely to present as an acute traumatic event.
- **No chronic lesion:** There is no well circumscribed subchondral defect with a loose body, which would suggest **osteochondritis dissecans** (option E); that condition is chronic/overuse related, not acute post traumatic.
- The film is clearly **not normal** (option A) given the visible fracture line and fragment.

Step 3 – Synthesis: Acute trauma + FOOSH mechanism + visible fracture at the radial head with preserved joint alignment = a fracture of the elbow (radial head fracture), not a dislocation or degenerative/chronic lesion.

Most likely diagnosis: C. Elbow fracture</think>

Most appropriate answer: \boxed{C}.