EVIDENCE-TO-CAPABILITY MATCHING FOR MEDICAL TRAINING DATA SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic training data for medical large language models (LLMs) draws on heterogeneous sources, but topical relevance does not guarantee that a source supports the capability a target task requires. We introduce **Evidence-to-Capability Matching** (ECM), a framework that profiles target capabilities, identifies source-supported evidence, and specifies requirements for constructing training questions. Its refinement component, **Medical Profile-Routed Agentic Refinement** (MEDPAR), coordinates source-constrained revision and verification for medical validity and capability alignment. On Qwen3-8B, task-focused corpora improve their primary benchmarks for evidence-intensive reasoning, clinical decisions, and knowledge coverage by 2.69, 6.68, and 1.46 percentage points, respectively. Ablations of six guidance families show that full guidance performs best in both tested settings, while the most consequential guidance differs by task: knowledge targeting for clinical decisions and evidence structuring for evidence-intensive reasoning. Clinical-decision-oriented training also improves on a benchmark excluded from synthesis and training, and a cross-task mixture improves over the base model on all four benchmarks. These results suggest that effective medical supervision depends on matching what a task requires to what its sources can support.

1 INTRODUCTION

Synthetic data offers a scalable way to turn existing text into post-training supervision for language models (Li et al., 2024; Nayak et al., 2024; Xu et al., 2025). Prior work emphasized quality and diversity, task-conditioned generation, and targeted selection (Liu et al., 2024; Nayak et al., 2024; Xia et al., 2024). Recent approaches further organize synthesis around explicit skills (Kaur et al., 2025) and student feedback (Khan et al., 2025), making learning needs an explicit input to data creation. These approaches help determine which examples are worth training on. In medicine, however, topical relevance and example quality leave a further question unresolved: *which properties of a source-grounded example support the medical capability we want to improve?* Answering this question requires specifying not only used source, but what supervision its evidence can justify.

This challenge arises because medical sources and target capabilities vary along different axes. A Case Report documents observations and decisions in a particular patient, whereas an examination question encodes a tested proposition and the distinctions needed to answer it. Guidelines and textbooks provide other forms of support, including recommendations and canonical knowledge (Wallace et al., 2022; Guyatt et al., 2008). Meanwhile, recalling a mechanism, selecting a clinical action, and integrating clues for a differential diagnosis demand different evidence and reasoning. A disease match alone cannot resolve whether a source supports the intended task: a rare treatment exception, for example, should not be rewritten as a general first-line recommendation.

The resulting gap is between *medical validity* and *task-dependent utility*. An example may be correct, source-supported, and answerable without exercising the intended capability; it may omit decisive clues, test a different clinical operation, or focus on an unhelpfully narrow distinction. Conversely, examples aimed at different tasks may share useful evidence or decision requirements. Thus, constructing useful supervision requires two linked decisions: what the learner should practice, and which parts of that requirement the source can actually support.

We address this gap with **Evidence-to-Capability Matching** (ECM; Figure 1). Rather than treating source selection and example generation as separate steps, ECM represents the target capability as

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.

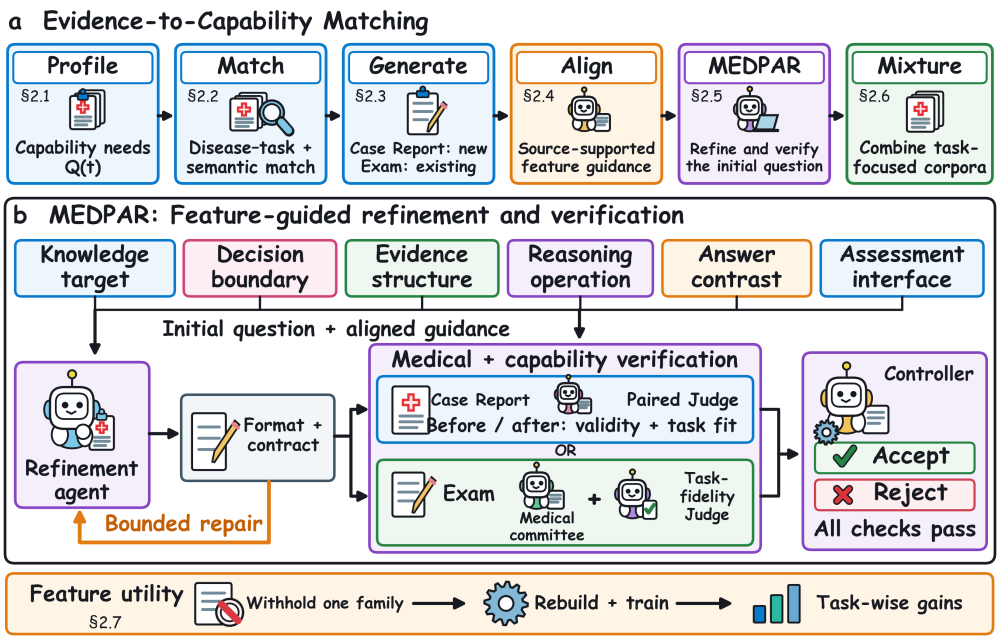

Figure 1: **ECM from profiles to training data.** (a) The pipeline connects target profiling, source matching, draft generation, source-supported alignment, refinement, and corpus assembly. GENERATE creates a Case Report draft or reuses an Exam question; ALIGN supplies source-supported requirements for refinement. (b) MEDPAR receives the initial question and these requirements, then refines and verifies the question. Bounded format and contract repair precedes final Judge decisions. The bottom strip illustrates the guidance-ablation study in Section 3.5.

an explicit profile and checks its requirements against source-supported evidence before construction. Its central component, **Medical Profile-Routed Agentic Refinement** (MEDPAR), converts the supported requirements into source-specific refinement and verification contracts. Specialized agents revise examples and check medical validity and capability alignment, while a rule-based controller coordinates repair. The method distinction is that the profile specifies *what the example should teach*, whereas source alignment limits *what the example may claim*; refinement must satisfy both.

This explicit interface also makes construction choices examinable. We organize the requirements into six guidance families and intervene on them during synthesis while preserving source allocation and medical-validity checks. The resulting comparisons probe which kinds of guidance matter for different target capabilities, rather than assuming that all valid examples have equal training utility. Experiments on Qwen3-8B show gains on the primary evaluations of three task-focused corpora; guidance ablations, non-target evaluation, and cross-task mixture training further examine how these gains depend on construction choices and extend beyond a single target. Our contributions are:

- We formulate medical synthetic-data construction as *evidence-to-capability matching*, separating the capability an example is intended to teach from the claims its source can support.
- We develop ECM and its MEDPAR component to turn this formulation into source-constrained, capability-directed refinement and verification.
- We introduce six interpretable guidance families as both construction requirements and intervention targets, and evaluate their task-dependent effects alongside non-target and cross-task performance.

2 METHOD

Overview. ECM constructs source-grounded supervision for a specified medical capability in two stages (Figure 1). Draft synthesis characterizes the target capability, retrieves and checks a medical source, and produces an initial question under requirements that the source can support. MEDPAR then revises the draft and verifies its medical validity and capability alignment. Accepted questions

form the corpora used in task-focused and cross-task training. We also intervene on the construction guidance to examine its contribution to downstream learning.

Problem Formulation. Let M_0 be an initial LLM, $\mathcal{S}$ a collection of medical sources, and $\mathcal{T}$ a set of target capabilities. For each $t \in \mathcal{T}$, we construct a corpus $\mathcal{D}_t$ of medical single-best-answer (SBA) questions. Each question contains a stem, answer options, a reference answer, and a bound source. Construction must satisfy two distinct requirements: the clinical claims and reference answer must be supported by the bound source under the permitted edits, and answering the question must require the operation specified by t. For example, a medically valid diagnosis question does not provide the intended supervision when the target capability is investigation selection.

After training $M_t = \mathrm{Train}(M_0, \mathcal{D}_t)$, we measure the effect on evaluation task u as

$$\Delta_{t \to u} = P_u(M_t) - P_u(M_0), \tag{1}$$

where $P_u(M)$ is accuracy on u. Source and capability checks determine whether a question is suitable for inclusion; downstream accuracy measures the utility of the resulting corpus rather than that of an individual question.

2.1 Source-Grounded Draft Synthesis

STEP-1 (PROFILE): Target Capability Guidance. Disease and clinical-task labels identify the subject of an assessment, but do not fully specify the capability needed to answer it. For a target capability t, PROFILE uses reference assessments to construct a target profile $Q(t)$ while retaining their disease/subject and clinical-task labels. For each reference question, the profiling prompt examines the stem, options, and available reference answer to identify what is being assessed, which evidence is needed, and how that evidence must be used. It also records presentation requirements and fixed constraints, such as option count and question direction.

The construction requirements in $Q(t)$ are organized into six guidance families, denoted by $\mathcal{F}_t$:

Knowledge Target: medical concept, learning objective, and clinical operation to be assessed.
Decision Boundary: patient conditions, exceptions, and priorities for judgment or action.
Evidence Structure: clinical clues needed and how they are presented and integrated.
Reasoning Operation: inference required to move from evidence to the answer.
Answer Contrast: distinctions between the correct answer and plausible alternatives.
Assessment Interface: vignette setting, level of detail, and question format.

These families describe complementary aspects of question construction, rather than separate task categories. They specify what a generated question should assess; subsequent source matching determines which requirements the available evidence can support. The same families serve as intervention units in Section 3.5. Detailed fields are provided in Appendix F.

STEP-2 (MATCH): Source Retrieval and Support Checks. The source pools contain 104,264 English Case Reports and 1,198,731 Exam records before filtering. During corpus preparation, we retain source identifiers and original text, exclude unusable or image-dependent material, and require Exam records to be parseable single-answer questions. Disease/subject and clinical-task labels are assigned at this stage and retained for matching.

Let $(q_{\mathrm{topic}}, q_{\mathrm{task}})$ denote the labels associated with target t, and let $L(s)$ be the label pairs assigned to source s. We restrict retrieval to sources with a matching label pair:

$$\mathcal{S}_t = \{s \in \mathcal{S} : (q_{\mathrm{topic}}, q_{\mathrm{task}}) \in L(s)\}. \tag{2}$$

Within $\mathcal{S}_t$, BGE embeddings Xiao et al. (2024) rank sources by semantic relevance to the target profile. Case Reports are ranked by their best-matching passage using task-text and profile-feature similarities, whereas Exams are ranked at the question level. The embeddings represent textual retrieval views; the categorical labels determine candidate eligibility. Source allocation prioritizes targets with fewer candidates and avoids reuse. When a Case Report passage determines its rank, the full report remains available for subsequent review. Retrieval views and scoring rules are detailed in Appendix B.

A relevant source does not necessarily support the intended question. For each candidate Case Report, a source auditor binds the question to one patient and clinical phase, distinguishes patient

observations from discussion-level claims, and records admissible evidence, permissible question scope, and unsupported directions. We denote this support record by $A(s)$. A source–target pair is rejected if the available evidence cannot support the required clinical operation. In particular, findings observed after a decision point cannot be presented as information available at that point, and general statements cannot be recast as observations about the bound patient.

STEP-3 (GENERATE): Initial Question. GENERATE produces an initial item x^- before the detailed guidance in $\mathcal{F}_t$ is applied. For case reports, a writer constructs the stem, candidate, and reference answer using the admissible patient evidence in $A(s)$, the target's coarse disease/task requirements, and the permissible question scope. For Exams, x^- is the source question with its supplied answer, which remains subject to medical review. In both cases, x^- is a starting point for subsequent refinement rather than an accepted training example. The question draft also serves as a fixed baseline for paired evaluation; no initial item enters the training corpus without final verification.

STEP-4 (ALIGN): Source-Supported Requirements. ALIGN determines which requirements in the target profile can be applied to the retrieved source and initial item x^-. For Case Reports, it uses the source-support record $A(s)$ to assess the admissible patient evidence and question scope; the generated draft is not treated as additional clinical evidence. For Exams, it checks proposed changes against the source question and its tested proposition. Supported requirements are retained, and requirements that can be met without changing the source-supported clinical proposition are adapted. Unsupported optional requirements are dropped; the source–target pair is rejected if the required clinical operation cannot be supported. The result is a source-specific refinement contract:

$$\mathcal{C}_{s,t} = (\mathcal{E}_s, \widetilde{\mathcal{F}}_{s,t}, o_{s,t}), \tag{3}$$

where $\mathcal{E}_s$ is the admissible source evidence or tested proposition, $\widetilde{\mathcal{F}}_{s,t}$ the feasible capability requirements, and $o_{s,t}$ the permitted edits. The contract specifies what refinement may preserve, emphasize, or change; subsequent verification determines whether the revised question satisfies these constraints.

2.2 MEDPAR: CONTRACT-GUIDED REFINEMENT AND VERIFICATION

Draft synthesis provides MEDPAR with an initial item x^-, its bound source s, and the alignment contract $\mathcal{C}_{s,t} = (\mathcal{E}_s, \widetilde{\mathcal{F}}_{s,t}, o_{s,t})$. The contract specializes the target guidance $\mathcal{F}_t$ to what s can support: it fixes the admissible evidence, identifies feasible capability requirements, and bounds permissible edits. MEDPAR revises x^- under the contract and admits the result to $\mathcal{D}_t$ only after verification.

STEP-1 (REVISE): Contract-Constrained Drafting. A writer revises x^- according to $\widetilde{\mathcal{F}}_{s,t}$ and records the evidence preserved and substantive changes made. For Case Reports, the revision uses the patient evidence and question scope established by $A(s)$; the draft itself cannot supply additional clinical facts. For Exams, it follows the permitted transformation while preserving the source question's tested proposition. The target profile $Q(t)$ informs the transformation decision, but agents receive only the requirements relevant to their roles.

STEP-2 (VERIFY): Checks and Bounded Repair. A rule-based controller first checks required components, consistency with the reference answer, and compliance with source-specific edit restrictions. Candidates passing these checks undergo complementary medical and capability reviews. Medical review assesses clinical correctness, fidelity to $\mathcal{E}_s$ and $o_{s,t}$, and whether the answer is supported and unambiguous. Capability review assesses the realized operation, evidence use, reasoning, and answer distinctions against $\widetilde{\mathcal{F}}_{s,t}$. Failures produce structured mismatch reasons. Those correctable within $\mathcal{C}_{s,t}$ may trigger targeted revision, subject to a fixed attempt limit. A failure requiring evidence outside $\mathcal{E}_s$ or an edit outside $o_{s,t}$ instead returns the candidate to source review. Repair therefore cannot expand the source-supported scope of the contract.

STEP-3 (ADMIT): Source-Specific Acceptance. For Case Reports, a paired Judge compares an eligible revision with the fixed draft x^- under the same patient evidence and feasible requirements. Correctness and unambiguous answerability are mandatory; core alignment scores must meet their thresholds, and the revision's total score must strictly exceed the draft's. For Exams, two generator-model calls and one independent-verifier call reassess the reference answer and its premises. Medical

acceptance requires at least two supporting votes, including the independent verifier's. A separate task-fidelity Judge checks permitted edits, currentness, reasoning, evidence efficiency, and response quality. The item must pass both medical and task checks. These judgments are screening criteria, not independent proof of correctness. Procedures and thresholds are given in Appendix B.3.

Items passing the applicable acceptance rule undergo final answer and format checks, duplicate filtering, and MinHash-based decontamination before entering $\mathcal{D}_t$ (Appendix B). We retain their source bindings, applied guidance, edit records, and review outcomes. All agents operate during offline construction; downstream evaluation uses only the trained model.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Training and Evaluation. Our primary study initializes Qwen3-8B Yang et al. (2025) and uses GRPO Shao et al. (2024) with thinking mode disabled. Each 8B run has a budget of 500 updates on 16 GPUs, with 64 prompts per update, 8 rollouts per prompt, and learning rate 10^{-6}. We additionally evaluate the same single-corpus configurations on Qwen3-30B-A3B; feature removals and mixture analysis use Qwen3-8B. Each model is compared with its own starting-model baseline. Every selected 8B model answers the same four benchmarks using greedy decoding; the A3B table reports each corpus's designated evaluation. Training configurations and checkpoint selection appear in Appendix C.1. Training dynamics and checkpoint curves are analyzed in Section 3.6.

3.2 DOWNSTREAM TASK DEFINITION

We organize the experiments around three synthesis priorities. Each priority has a corresponding synthetic training corpus and primary evaluation benchmark:

(1) The ***Evidence-Intensive Synthesis (EI)*** corpus contains 4,925 Case Report–Exam examples, excluding those answered correctly in all five teacher rollouts. We use MedXpertQA-Text (2,450 questions) as its primary evaluation. This benchmark includes difficult licensing- and specialty-board questions, with information-rich clinical scenarios and ten answer options that demand careful discrimination among plausible alternatives (Zuo et al., 2025). These characteristics make it suitable for assessing evidence integration and complex clinical reasoning.

(2) The ***Clinical-Decision Synthesis (CD)*** corpus contains 6,539 refined Case Report examples. Its primary evaluation is the MedQA subset used in our experiments (1,273 questions). Its patient vignettes require interpreting clinical findings to identify a diagnosis, select an investigation, or determine an appropriate treatment (Jin et al., 2021). Such questions often involve linked reasoning steps, making this benchmark suitable for assessing patient-specific clinical judgments.

(3) The ***Knowledge-Coverage Synthesis (KC)*** corpus contains 12,000 Exam and Case Report examples, without a separate hard-reasoning replay component. We use MedMCQA's development split (4,183 questions) as its primary evaluation. Built from postgraduate medical entrance examinations, MedMCQA covers 21 medical subjects and approximately 2,400 topics, with questions involving conceptual understanding, option elimination, diagnosis, and treatment (Pal et al., 2022). This breadth makes it suitable for assessing knowledge coverage across medical subjects.

(4) HLE_verified (215 questions) evaluates ***Non-Target Generalization (NT)***. Its questions are excluded from synthesis and training, and no HLE-targeted corpus is constructed. NT denotes an evaluation direction, not a training corpus.

3.3 SCENARIO 1: TASK-FOCUSED CAPABILITY LEARNING

Setup. We train separate models on EI, CD, and KC corpora. Each model is evaluated on all benchmarks, with MedXpertQA, MedQA, and MedMCQA serving as the respective primary benchmarks. This scenario examines improvement on the intended capability as well as transfer to other tasks.

Results on Qwen3-3B Task-focused training yields different gains across benchmarks (Table 1). EI improves MedXpertQA from 16.53% to 19.22% (+2.69 points), CD improves MedQA from 71.72% to 78.40% (+6.68 points), and KC improves MedMCQA from 62.18% to 63.64% (+1.46 points). CD

Table 1: Accuracy (%) across the four evaluation benchmarks. EI, CD, and KC are task-focused conditions; CD→NT tests generalization to HLE_verified, which is excluded from synthesis and training. The cross-task condition uses a fixed 2:1:1:1 mixture. Parenthetical values highlight gains over the Qwen3-8B baseline on the indicated benchmarks.

Benchmark	Baseline	Task-Focused			Generalization	Cross-Task
		EI	CD	KC	CD→NT	2:1:1:1
N	Qwen3-8B	4,925	6,539	12,000	11,852	16,000
MedXpertQA	16.53	19.22(**+2.69**)	19.71	16.61	19.35	18.94
MedQA	71.72	71.33	78.40(**+6.68**)	74.08	73.45	74.55
MedMCQA	62.18	61.03	62.16	63.64(**+1.46**)	62.04	62.44
HLE-verified	24.19	22.79	25.12	24.65	26.98(**+2.79**)	26.05
Avg.	43.65	43.59	46.35	44.74	45.45	45.49

Cross Task Mixture Ratio: CD:ER:KC:NT=2:1:1:1.

Table 2: Qwen3-30B-A3B accuracy (%). Parentheses show gains over the baseline before rounding.

Benchmark	Condition	Baseline	Trained (gain)
MedXpertQA	EI	19.71	**21.67 (+1.96)**
MedQA	CD	79.73	**83.50 (+3.77)**
MedMCQA	KC	68.35	**69.59 (+1.24)**
HLE-verified	CD→NT	26.05	**27.44 (+1.40)**

also achieves the highest four-task average (46.35%) and the highest MedXpertQA-Text accuracy (19.71%), despite the latter not being its primary benchmark. In contrast, EI falls below the baseline on the other three benchmarks. Thus, primary-task gains do not necessarily imply broad transfer.

Results on Qwen3-30B-A3B. Table 2 examines whether the gains observed on Qwen3-8B persist on Qwen3-30B-A3B. All three task-focused conditions improve on their primary benchmarks: EI gains 1.96 points on MedXpertQA, CD gains 3.77 points on MedQA, and KC gains 1.24 points on MedMCQA. Clinical-decision-oriented training also improves the non-target HLE-verified benchmark by 1.40 points. Although the primary-task gains are smaller than those observed on Qwen3-8B, their consistent direction suggests that the benefit of capability-matched supervision extends to this larger model. Because the 30B experiments report selected condition–benchmark pairs rather than a full evaluation matrix, we do not infer broader cross-task transfer from these results.

3.4 Scenario 2: Cross-Task Mixture Generalization

Setup. To examine whether clinical-decision-oriented supervision generalizes beyond its target task, we evaluate an additional 11,852-example condition denoted CD→NT. The arrow identifies the generalization being tested: HLE_verified provides the non-target evaluation and is excluded from synthesis and training. We also train a 16,000-example cross-task mixture (CT) using a fixed 2:1:1:1 ratio by example count. The ratio is specified in advance rather than optimized using Judge scores; component counts and the corresponding mixture inputs appear in Appendix C.3. We evaluate both conditions on all four benchmarks.

Results. CD→NT achieves 26.98% on HLE_verified, 2.79 points above the Qwen3-8B baseline and 1.86 points above the task-focused CD model (Table 1). It obtains the highest HLE_verified accuracy among the compared conditions, providing evidence of generalization to this non-target benchmark without using its questions for training. CT exceeds the baseline on all four benchmarks and raises the four-task average from 43.65% to 45.49%. Its HLE_verified accuracy is 26.05%: above the baseline, but below CD→NT.

Table 3: **Guidance ablations.** Accuracy (%) on MedQA (CD) and MedXpertQA (EI); parentheses show changes from Full before rounding.

Condition	MedQA	MedXpertQA
Qwen3-8B baseline	71.72	16.53
Full six guidance	**78.40**	**19.22**
w/o Knowledge Target	72.58 (−5.81)	18.20 (−1.02)
w/o Decision Boundary	73.29 (−5.11)	18.57 (−0.65)
w/o Evidence Structure	75.18 (−3.22)	17.80 (−1.43)
w/o Reasoning Operation	73.53 (−4.87)	17.96 (−1.27)
w/o Answer Contrast	73.13 (−5.26)	18.49 (−0.73)
w/o Assessment Interface	73.06 (−5.34)	18.24 (−0.98)

3.5 Scenario 3: Utility of Capability Guidance

Setup. We test the six guidance families separately for clinical-decision and evidence-intensive synthesis. Within each setting, Full uses all six families. Each ablation withholds one family's instructions, including equivalent instructions and dependent fields, while reusing the same source records and stored guidance. No new retrieval or source replacement is performed; source facts and medical-validity checks remain in place. Withholding guidance makes the corresponding property unconstrained, not prohibited.

Under the same model initialization and training recipe, we compare each ablation with Full from the same synthesis setting:

$$\Delta_f(u \mid M_0) = P_u(M^*_{\text{Full}}) - P_u(M^*_{-f}), \tag{4}$$

where M^*_a is the selected checkpoint for condition a and P_u is accuracy on evaluation task u. A positive Δ_f favors Full. These contrasts estimate the contribution of providing guidance under this construction procedure; they do not establish that the corresponding property is absent from ablation data. Figure 6 illustrates the intervention, and Appendix E gives construction details.

Results over Different Benchmarks. Full guidance achieves the best accuracy in both settings (Table 3). In clinical-decision synthesis, withholding knowledge-target produces the largest accuracy drop (5.81 on MedQA). For evidence-intensive, the largest drops follow removal of evidence-structure (1.43) and reasoning-operation guidance (1.27) on MedXpertQA. The ordering of these effects differs across settings, consistent with task-dependent guidance utility. In each ablation, we remove one guidance, not the corresponding content from generated questions. The results therefore measure the value of giving that guidance during construction.

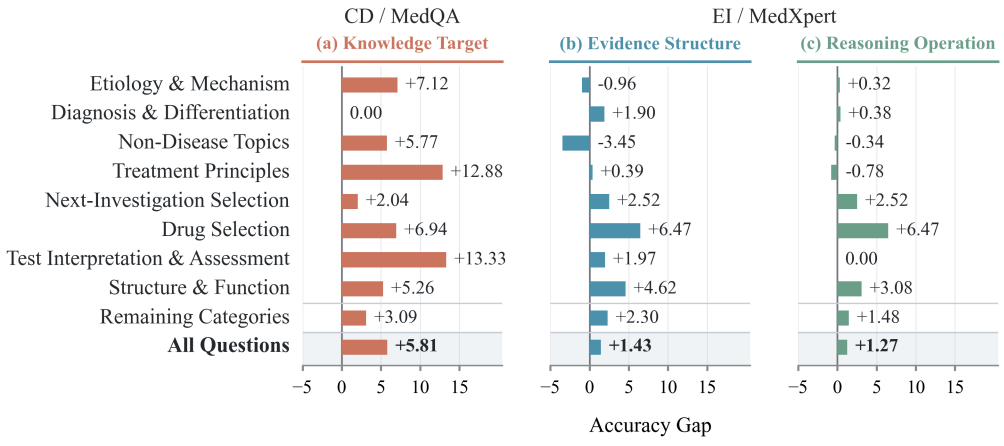

Figure 2: **Guidance effects across clinical tasks.** This plot illustrates the contribution from specific guidance on different clinical tasks by showing the accuracy gap between the full guidance set and the group without a spefific guidance. Category definitions and counts appear in Appendix G.1.

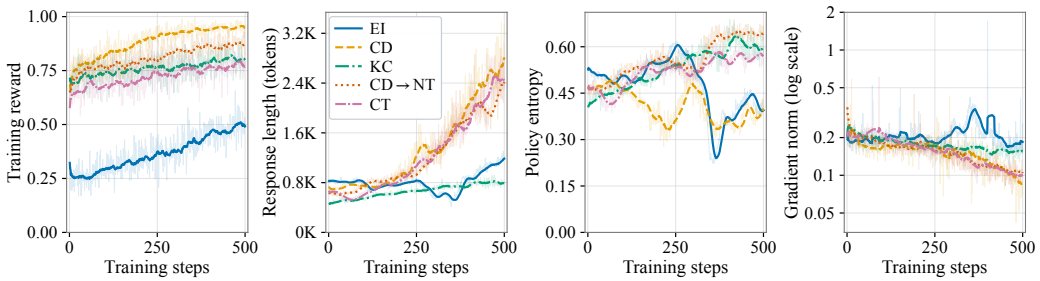

Figure 3: **Training dynamics of Qwen3-8B over 500 GRPO updates.** Faint curves show per-update values; dark curves show trailing 20-update means.

Results over Different Clinical Tasks. To see where the aggregate ablation gaps arise, we group evaluation questions by their existing clinical-task labels. These labels describe what an answer must accomplish—for example, explaining a mechanism, interpreting findings, or selecting treatment—rather than its disease topic or source genre. This connects the analysis to ECM's capability goals without treating each benchmark as a uniform skill. The labels are fixed across models; they categorize evaluation questions, whereas the removed guidance governs training-question construction.

Figure 2 expands the largest aggregate contrasts in Table 3. In CD, full knowledge-target guidance improves treatment-principle and test-interpretation accuracy by 12.88 and 13.33 percentage points, respectively, while diagnosis totals are unchanged. Its benefit therefore extends beyond disease recognition to the relations and rules a clinical question tests. In EI, full evidence-structure and reasoning-operation guidance each improve drug selection by 6.47 points and next-investigation selection by 2.52 points. Evidence organization and inference requirements thus also matter for action selection, not only for a separate class of reasoning questions. These descriptive comparisons locate the aggregate differences within clinical operations; they do not assign individual errors to a single missing feature.

3.6 Training Dynamics and Checkpoint Analysis

Training Dynamics. Figure 3 shows increasing training reward across all five conditions, accompanied by different response behaviors. CD, CD→NT, and CT produce substantially longer responses, averaging approximately 2,200–2,700 tokens over the final 50 updates, compared with 800–1,100 for KC and EI. Reward gains therefore accompany different degrees of response expansion. Since the corpora differ, absolute reward levels do not rank the conditions, and longer responses alone do not establish better reasoning. Policy entropy reveals a different grouping: it generally rises for KC, CD→NT, and CT, while CD and EI exhibit pronounced declines and recoveries. In particular, CD produces longer responses despite lower mean entropy, whereas KC remains relatively short while entropy rises. Gradient norms decrease more strongly for CD, CD→NT, and CT, more moderately for KC, and show transient spikes without a comparable sustained decline for EI. Together, the curves show that reward improvement does not correspond to one shared pattern of length, entropy, or gradient change. These single-run observations describe training behavior; downstream benefits are assessed separately through benchmark evaluation.

Cross-Task Performance During Training. To examine how cross-task performance evolves, we compare CT at ten training checkpoints (Figure 4). From step 320 to 500, MedXpertQA improves from 18.94% to 19.47%, while MedQA decreases from 74.55% to 72.27% and MedMCQA from 62.44% to 61.87%. HLE-verified fluctuates more sharply: it reaches 26.98% at step 440, falls to 21.86% at 480, and ends at 23.72%. The four-task mean consequently decreases from 45.49% at step 320 to 44.33% at step 500, despite rising training reward. These trajectories show that continued optimization produces different gains and losses across tasks, rather than a uniform improvement in downstream performance.

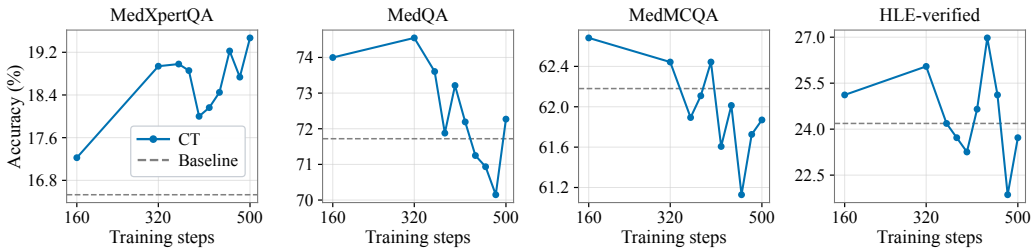

Figure 4: **CT performance across training checkpoints.** Accuracy on four benchmarks; dashed lines show the Qwen3-8B baseline.

3.7 Analysis: What Makes Capability Guidance Useful?

The results suggest that useful guidance resolves two ambiguities left by source relevance: *which medical relation to teach* and *which inference the learner must perform*. In CD, knowledge-target removal produces the largest loss despite reusing the source records. The gap appears in treatment principles and test interpretation, while diagnosis totals are unchanged (Figure 2). A disease match therefore does not fully specify the learning objective: a question can concern the right condition yet fail to practice the intended relation or decision rule. The CD construction example illustrates this distinction by replacing an unsupported mechanism question with a source-supported hemodynamic target (Appendix G.2).

In EI, evidence structure and reasoning operation matter most in the observed comparisons. Stating an intermediate diagnosis can bypass the intended inference even when every fact is correct; the EI example instead retains the patient's clues without supplying that conclusion. Refinement thus governs not just the amount of evidence, but what work remains for the learner.

These priorities are complementary, not a division into knowledge-only and reasoning-only tasks. Evidence and reasoning guidance also benefit drug and investigation selection, while knowledge targeting supports interpreting findings and choosing treatment. This connects the stages of ECM: Profile specifies the intended operation, Match supplies relevant sources, and Align/MedPAR translate their admissible evidence into questions that require that operation. Full remains best in both studies, so prioritizing one guidance family does not justify dropping the others. The ablations therefore serve two purposes: they assess the synthesis system and identify what its reviewers should inspect most closely for each task, providing a task-specific interpretation of otherwise opaque data-quality judgments. The A3B results extend corpus utility to another backbone, whereas the guidance rankings remain conditional on the tested Qwen3-8B construction settings.

4 Conclusion

We presented ECM, a framework for constructing source-grounded medical supervision for specified clinical capabilities. The method separates source relevance from clinical support, translates target capability guidance into source-specific refinement contracts, and admits revised items only after medical and capability verification. This separation is important because a clinically valid question need not assess the intended operation, while a capability-aligned question may still rely on claims its source does not support. By retaining source bindings, feasible requirements, and review outcomes, the framework makes the construction process auditable. The task-focused, cross-task, and guidance-removal experiments examine whether the resulting supervision improves downstream performance and which construction requirements contribute to that effect. More broadly, the framework offers a way to study medical question construction as constrained capability alignment rather than unrestricted question generation.

AI USE STATEMENT

In preparing the manuscript, we used AI tools solely for language polishing, including grammar, phrasing, and readability. The research questions, methodological choices, and experimental protocols were determined by the authors. The authors reviewed and approved all AI-assisted edits to ensure that they preserved the intended technical meaning, and take full responsibility for the final manuscript, including its claims and conclusions.

REPRODUCIBILITY STATEMENT

The paper documents the method and experimental protocols to support reproducibility. Section 2 defines the target profiles, source-matching procedure, synthesis constraints, and refinement controller. Appendix B provides source-specific algorithms, retrieval settings, acceptance criteria, and filtering procedures. Section 3 describes the starting model, training settings, datasets, and evaluation setup; Appendix C specifies scoring, checkpoint selection, and mixture construction. Appendix E details the feature-removal protocol and additional results, with the guidance fields enumerated in Appendix F. The paired-analysis procedure, resampling seed, and multiple-comparison correction are given in Appendix E.4. The evaluation protocol also states the limits of single-trajectory results and benchmark-based checkpoint selection, so that reproduction and interpretation use the same comparison conditions.

ETHICS STATEMENT

This study uses medical source materials and benchmark datasets for offline research on training-data synthesis and medical question answering. The data were used in compliance with applicable data-use requirements and source permissions, with respect for patient privacy and the permitted research scope of the source materials. Our aim is to improve the construction and understanding of source-grounded supervision for medical language models. The source-grounding constraints and verification procedures described in this paper support fidelity to the underlying medical evidence and the quality of the synthesized supervision. We report our findings within the scope of the evaluated tasks and experimental settings.

REFERENCES

Gordon H Guyatt, Andrew D Oxman, Gunn E Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, and Holger J Schünemann. Grade: an emerging consensus on rating quality of evidence and strength of recommendations. *bmj*, 336(7650):924–926, 2008.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021. ISSN 2076-3417. URL https://www.mdpi.com/2076-3417/11/14/6421.

Simran Kaur, Simon Park, Anirudh Goyal, and Sanjeev Arora. Instruct-skillmix: A powerful pipeline for llm instruction tuning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 28842–28983, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/473a9a75edc46eff5ff224d53d5f7294-Paper-Conference.pdf.

Zaid Khan, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Dataenvgym: Data generation agents in teacher environments with student feedback, 2025. URL https://arxiv.org/abs/2410.06215.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, and Mike Lewis. Self-alignment with instruction backtranslation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 3552–3577, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0f8e3534eb8dee7478d4dc0e9d9a0b1a-Paper-Conference.pdf.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 22353–22373, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6091f2bb355e960600f62566ac0e2862-Paper-Conference.pdf.

Nihal Nayak, Yiyang Nan, Avi Trost, and Stephen Bach. Learning to generate instruction tuning datasets for zero-shot task adaptation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12585–12611, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.748. URL https://aclanthology.org/2024.findings-acl.748/.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Sowdhamini S Wallace, Gal Barak, Grace Truong, and Michelle W Parker. Hierarchy of evidence within the medical literature. *Hospital pediatrics*, 12(8):745–750, 2022.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings, 2024. URL https://arxiv.org/abs/2309.07597.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 76346–76382, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/be06e3802e9411381feece79b4d960c1-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80961–80990. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zuo25a.html.

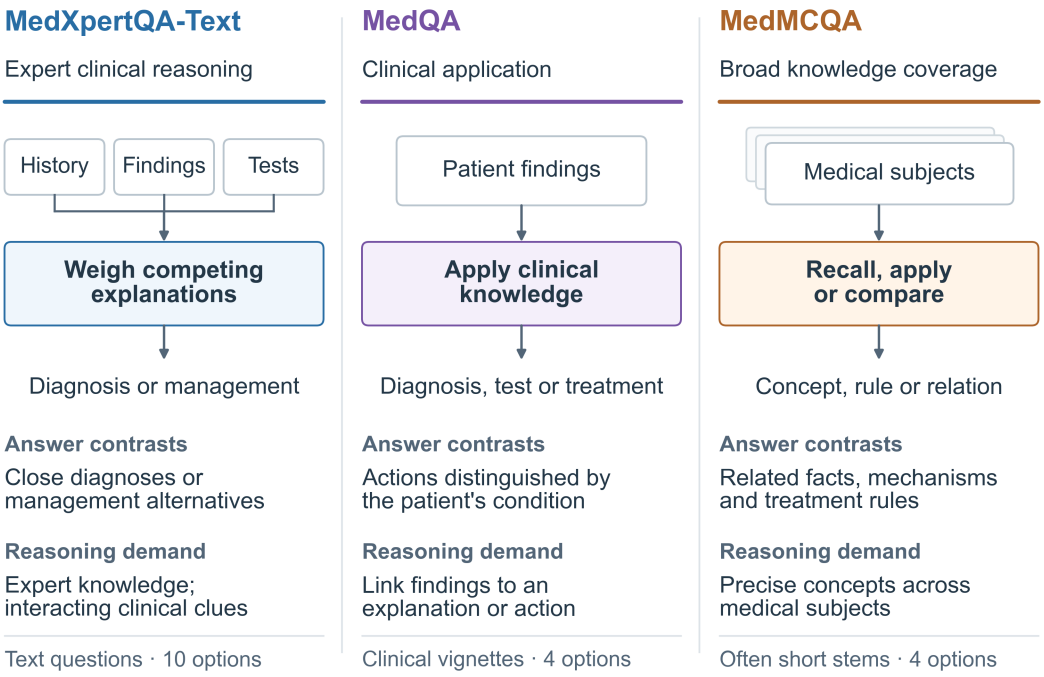

Figure 5: **Contrasting task-family profiles.** The shared schema describes what is assessed, how evidence is used, which answers must be distinguished, and the reasoning demand. The figure summarizes benchmark designs (Zuo et al., 2025; Jin et al., 2021; Pal et al., 2022) and task-style requirements, not individual questions or measured label frequencies. These are overlapping emphases, not fixed properties of every item.

A Task-Family Profiles

The shared profiling schema describes the learning target, evidence use, answer contrasts, and reasoning demand, with interface requirements recorded alongside them. Figure 5 summarizes how these dimensions differ across the three evaluation families; it is a qualitative comparison of their documented emphases, not a claim that every question in a family has the same structure.

B Source-Specific Synthesis Details

B.1 Retrieval and routing

Equation 2 uses existing disease–task annotations for candidate filtering. For semantic ranking, let $e(\cdot)$ be the BGE text encoder, q_t the reference task-example text associated with profile $Q(t)$, and f_t the textual serialization of its retrieval features. $\mathrm{chunks}(s)$ denotes passages extracted from source s, and $\mathbb{I}_{\mathrm{primary}}$ equals one when the matched task is the source's primary label and zero otherwise. Case Reports use the score

$$S_{\mathrm{case}}(s,t) = \max_{c \in \mathrm{chunks}(s)} \left[\lambda_q \cos(e(c), e(q_t)) + \lambda_f \cos(e(c), e(f_t))\right] + \delta_p \mathbb{I}_{\mathrm{primary}}. \tag{5}$$

We fix $\lambda_q = 0.55$, $\lambda_f = 0.45$, and $\delta_p = 0.025$, giving topical semantics slightly more weight than profile features and a small preference to primary task labels. Exam sources use $\cos(e(s), e(q_t))$ after the same disease–task restriction, retaining similarities in $[0.45, 0.82)$ in the tiered configuration.

Reference assessment examples supply inputs to profile extraction and the task-text retrieval view. Existing disease and task assignments are reused rather than re-labeled. The retrieved source is then audited, and synthesis follows the source-supported contract in Equation 3. Thus the ranking stage and the generation stage use distinct input representations.

For Case Reports, deterministic rules reject wrong-disease, non-human, non-patient, or insufficient-evidence sources and otherwise select evidence-preserving adaptation or source-grounded recomposition. For Exam sources, independent medical votes certify the proposition and answer before the route grants proposition preservation, reasoning hardening, local repair, topic-level reconstruction, or rejection.

B.2 Quality and difficulty filtering

Source-specific verification follows the acceptance rules below. Final corpora undergo format checks, answer verification, and MiniHash decontamination. Teacher responses, when produced, are screened for answer agreement and output validity. The evidence-intensive corpus additionally uses DeepSeek-V4-Flash Pass@5: examples answered correctly in all five rollouts are excluded, retaining those solved in zero to four rollouts. This release-specific difficulty filter is not applied to every corpus.

B.3 Source-specific synthesis algorithms

The algorithms expand Figure 1 using shared profiles $Q(t)$, source labels $L(s)$, and retrieval views q_t, f_t. Match, Generate, and Align correspond to Sections **??–??**; refinement and verification implement MedPAR (Section 2.2). Source eligibility is checked before a new Case Report draft; the resulting support record is reused for feature projection. Let $\mathcal{R}_t$ be a ranked candidate list, k the source quota per profile, and $\mathcal{P}$ the allocated source–profile pairs. Each algorithm returns accepted questions and answer labels in $\mathcal{D}$.

Case Report: retrieve passages, but generate from patient-bound evidence. In Algorithm 1, $v(s)$ is the report text or an available task-specific evidence plan/abstract. ParagraphChunks packs paragraphs up to $\ell = 1800$ characters; oversized paragraphs use windows with overlap $\omega = 300$. At most $m = 12$ chunks are retained in source order. MaxChunkScore applies Equation 5 and selects the best-matching chunk c^*. UniqueAllocate prioritizes profiles with fewer candidates, then allocates unused sources round-robin up to quota k. Generation receives document context capped at 30,000 characters, not only the retrieved chunk.

The source auditor returns the patient binding p_s, admissible facts $\mathcal{E}_s$, and the support profile $A(s)$, including phase, allowed question scope, forbidden directions, and per-feature support. CaseRoute selects direct construction, evidence-preserving adaptation, source-grounded structural recomposition, or rejection. Recomposition runs only when that route is enabled. ProjectFeatures retains supported requirements, substitutes source-consistent adaptations, and drops unsupported requests; recomposition transfers only compatible abstract structure. The initial question x^- uses the patient evidence and coarse task requirements, while x^+ additionally realizes the selected feature contract. PairedAccept requires correctness, a unique answer, no flagged near-copy, core scores of at least $\tau_c = 3$ on the 0–4 scale, and a strictly greater total score than x^-. The core scores cover disease, task, answer type, reasoning, option design, difficulty, and source fidelity; question structure also contributes to the total. The rewrite budget is $R_c = 3$ attempts.

Exam: verify the proposition before deciding how far to rewrite. Algorithm 2 ranks complete, text-only single-answer questions without chunk aggregation, using the same source-uniqueness allocation rule. Its similarity bounds are $\sigma_{\min} = 0.45$ and $\sigma_{\max} = 0.82$. MedicalCommittee independently solves and checks an item with two generator-model votes and one independent-verifier vote. Acceptance requires at least two votes supporting a correct, unique answer, including the independent verifier. This rule applies both before and after rewriting.

An audited source that passes can be preserved or strengthened around the same answer; a failed source can be repaired, reconstructed around a useful topic, or rejected. ExamContract specifies the permitted changes and task-level style requirements. Preservation forbids new patient facts; hardening permits minimal standard background without changing the answer concept; repair addresses identified defects; reconstruction may change the proposition and answer. RouteAndTaskJudge checks these permissions, medical currentness, reasoning, evidence efficiency, and option quality. Both high-fit and valid-standard acceptance tiers require correctness and a unique answer. The generation budget is $R_e = 2$; U denotes the failed checks supplied to a rewrite.

Algorithm 1 Case Report: chunk-based matching, patient-bound generation, and paired refinement

Input: Labeled reports $\mathcal{S}_c$; profiles $Q(t)$ and retrieval views (q_t, f_t)
Input: Source quota k; chunk parameters (ℓ, ω, m); routes $\mathcal{O}_c$; attempt budget R_c
Output: Accepted questions and answer labels $\mathcal{D}_c$

$\mathcal{D}_c \leftarrow \varnothing$
Match: coarse labels, then passage-level semantic ranking
for all requested profiles $t \in \mathcal{T}$ **do**
  $\mathcal{R}_t \leftarrow \varnothing$; $\mathcal{S}_t \leftarrow \textsc{MatchLabels}(\mathcal{S}_c, L, Q(t))$
  for all $s \in \mathcal{S}_t$ **do**
    $\mathcal{B}_s \leftarrow \textsc{ParagraphChunks}(v(s), \ell, \omega, m)$
    $(r, c^*) \leftarrow \textsc{MaxChunkScore}(\mathcal{B}_s, q_t, f_t, L(s))$ ▷ Eq. 5
    Add (s, t, r, c^*) to $\mathcal{R}_t$
  end for
end for
$\mathcal{P} \leftarrow \textsc{UniqueAllocate}(\{\mathcal{R}_t\}, k)$ ▷ No source reuse
for all allocated pairs $(s, t, c^*) \in \mathcal{P}$ **do**
  Source preparation: bind patient evidence and cache support decisions
  $u_s \leftarrow \textsc{SourceContext}(s)$
  $(A(s), p_s, \mathcal{E}_s) \leftarrow \textsc{SourceAudit}(u_s, Q(t))$
  $o \leftarrow \textsc{CaseRoute}(A(s), Q(t))$
  if o is rejected or not in $\mathcal{O}_c$: **continue**
  Generate: establish a source-grounded initial question
  $x^- \leftarrow \textsc{InitialWriter}(u_s, p_s, \mathcal{E}_s, \text{coarse task/interface})$
  if x^- is invalid: **continue**
  Align: form source-supported guidance for refinement
  $\widetilde{\mathcal{F}}_{s,t} \leftarrow \textsc{ProjectFeatures}(Q(t), A(s), o)$ ▷ Keep/adapt/drop
  $\mathcal{C}_{s,t} \leftarrow (\mathcal{E}_s, \widetilde{\mathcal{F}}_{s,t}, o)$ ▷ Eq. 3
  MEDPAR: refine the initial question, then verify the revision
  $P \leftarrow \textsc{RefinementPrompt}(u_s, x^-, \mathcal{C}_{s,t})$; $h \leftarrow \text{false}$
  for $j = 1, \ldots, R_c$ **do**
    $x^+ \leftarrow \textsc{Writer}(P)$
    $(h, U) \leftarrow \textsc{CheckStructureAndContract}(x^+, \mathcal{C}_{s,t})$
    if h or the candidate is not repairable: **break**
    $P \leftarrow \textsc{AddRepairInstructions}(P, U)$ ▷ Specific failed checks
  end for
  if $\neg h$: **continue**
  $J \leftarrow \textsc{PairedJudge}(u_s, \mathcal{C}_{s,t}, x^-, x^+)$
  if $\textsc{PairedAccept}(J, \tau_c)$ **then**
    Add x^+ and its answer to $\mathcal{D}_c$
  end if
end for
$\mathcal{D}_c \leftarrow \textsc{CorpusChecks}(\mathcal{D}_c)$ ▷ Format, duplicates, quotas
return $\mathcal{D}_c$

Algorithm 2 Exam: question-level matching, medically audited routing, and dual verification

Input: Labeled source exams $\mathcal{S}_e$; profiles $Q(t)$ and retrieval text q_t
Input: Source quota k; similarity interval $[\sigma_{\min}, \sigma_{\max})$; attempt budget R_e
Output: Accepted questions and answer labels $\mathcal{D}_e$

$\mathcal{D}_e \leftarrow \varnothing$
$\mathcal{S}_e \leftarrow \textsc{ParseAndFilter}(\mathcal{S}_e)$ ▷ Text-only, complete single-answer items
Match: coarse labels, then question-level semantic ranking
for all requested profiles $t \in \mathcal{T}$ **do**
  $\mathcal{R}_t \leftarrow \varnothing$; $\mathcal{S}_t \leftarrow \textsc{MatchLabels}(\mathcal{S}_e, L, Q(t))$
  for all $s \in \mathcal{S}_t$ **do**
    $r \leftarrow \cos(e(\text{question}(s)), e(q_t))$
    if $\sigma_{\min} \le r < \sigma_{\max}$: add (s, t, r) to $\mathcal{R}_t$
  end for
end for
$\mathcal{P} \leftarrow \textsc{UniqueAllocate}(\{\mathcal{R}_t\}, k)$ ▷ No source reuse
for all allocated pairs $(s, t) \in \mathcal{P}$ **do**
  Generate: recover the existing source question as x^- ▷ No new draft
  $x^- \leftarrow \text{question_and_answer}(s)$
  Align: audit source support and determine rewrite permissions
  $V_s \leftarrow \textsc{MedicalCommittee}(s)$
  $B \leftarrow \textsc{SourceAndProfileAudit}(s, V_s, Q(t))$
  if V_s passes **then**
    $o \leftarrow$ preserve if operation/answer type match; otherwise harden
  else if B identifies a localized repairable defect **then**
    $o \leftarrow$ local repair
  else if B supports retaining a useful topic **then**
    $o \leftarrow$ topic reconstruction
  else
    continue
  end if
  if (o, B, V_s) is inconsistent: **continue**
  $\mathcal{C}_{s,t} \leftarrow \textsc{ExamContract}(s, B, o, \text{task-level style})$
  MEDPAR: revise under the route and repair validation failures
  $U \leftarrow \varnothing$; $h \leftarrow \text{false}$
  for $j = 1, \ldots, R_e$ **do**
    $x^+ \leftarrow \textsc{RouteWriter}(s, \mathcal{C}_{s,t}, U)$
    $(h, U) \leftarrow \textsc{CheckFormatAndPermissions}(x^+, s, o)$
    if h: **break**
  end for
  if $\neg h$: **continue**
  Rebalance answer positions when the route permits changing the answer
  Verify: independently test medicine and route/task fidelity
  $V_x \leftarrow \textsc{MedicalCommittee}(x^+)$ ▷ Fresh votes on the new item
  $J \leftarrow \textsc{RouteAndTaskJudge}(s, x^+, \mathcal{C}_{s,t})$; $a \leftarrow \textsc{Tier}(J)$
  if V_x passes, a is accepted, and $\textsc{SimilarityCheck}(x^+, t)$ passes **then**
    Add $(x^+, \text{answer}(x^+), a)$ to $\mathcal{D}_e$
  end if
end for
return $\mathcal{D}_e$

C TRAINING, EVALUATION, AND DATA MIXTURES

C.1 TRAINING CONFIGURATION

We use the same GRPO optimization settings across the training corpora and feature-removal conditions, as summarized in Table 4. Training is implemented in `verl`, with fully sharded data parallelism for policy updates and `vLLM` for rollout generation. Each run uses two nodes with eight GPUs per node, tensor parallelism of four for generation, and sequence parallelism of four for policy updates. Gradient checkpointing and dynamic microbatching are enabled.

Training rewards use the extracted final answer: a correct option receives 1 and an incorrect or unparseable answer receives 0. A repetition guard assigns -1 when the same boxed answer appears at least ten times in the final response. No additional length penalty is applied. Training rollouts are sampled at temperature 1; the reported benchmark results use the separate greedy evaluation protocol below.

Table 4: **GRPO training hyperparameters.** The prompt limit is 1,152 tokens for the main runs and clinical-decision removals, and 1,280 tokens for evidence-intensive removals.

Hyperparameter	**Value**
Training budget	500 updates
Prompts per update / rollouts per prompt	64 / 8 (512 responses)
Policy minibatch size / epochs per update	64 / 1
Learning rate / schedule / warmup	10^{-6} / constant / 0 steps
Weight decay / gradient clipping norm	0.01 / 1.0
Lower / upper policy clipping parameters	0.20 / 0.28
KL regularization / entropy bonus	Disabled / disabled
Rollout temperature / top-p / top-k	1.0 / 1.0 / disabled
Maximum training response length	18,848 tokens
Maximum generation context length	40,960 tokens
Rollout precision	`bfloat16`
Checkpoint interval / random seed	20 updates / 42

C.2 EVALUATION AND REPORTING

All models use the benchmark sets in Section 3, with greedy decoding at temperature zero and one response per question. MedXpert uses the answer-rule scorer; the other benchmarks use the GPT-4o scoring configuration. Table 1 reports accuracies and their unweighted four-task mean, with differences computed before rounding.

Checkpoints are selected by the primary evaluation for EI, CD, and KC, by HLE_verified accuracy for ER, and by the four-task mean for CT. The selected steps are 360, 420, 360, 420, and 320, respectively. HLE_verified questions are excluded from synthesis and training, but the benchmark informs selection for ER and CT; it is not an independent model-selection holdout. Results come from single training trajectories; corpus and mixture differences are reported as point estimates rather than statistically established rankings.

ER contains 11,852 examples from completed Exam items and decontaminated MedQA training questions, audited and retained or revised under a shared clinical-exam style profile. It uses no new per-profile source retrieval and is an additional synthesis configuration, not a controlled retrieval ablation.

C.3 CROSS-TASK MIXTURE COMPOSITION

The cross-task mixture (CT) combines 16,000 examples from completed capability-specific corpora: 6,000 clinical-decision synthesis examples (37.5%), 4,000 Exam-based refinement examples (25%), 3,200 knowledge-coverage synthesis examples (20%), and 2,800 evidence-intensive synthesis examples (17.5%). Proportions are defined by example count. Clinical-decision supervision receives the largest share, while the other components provide complementary knowledge coverage, examination-

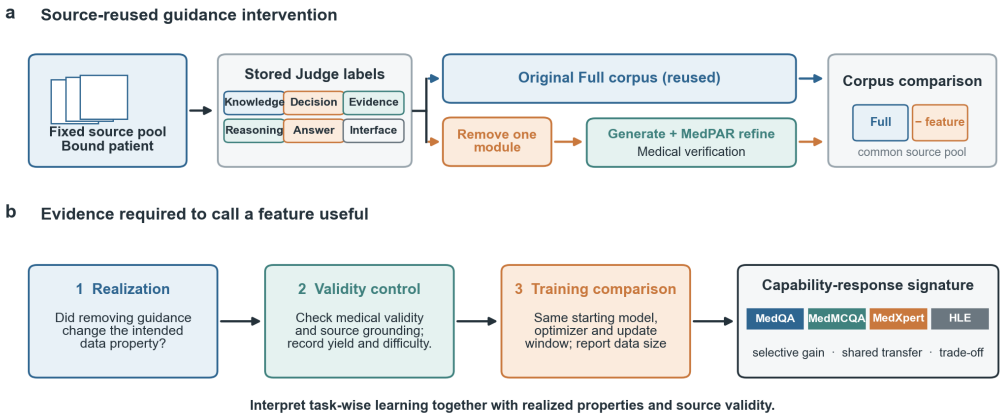

Figure 6: **From guidance intervention to learning utility.** Full retains the original corpus; each removal withholds one guidance family when constructing questions from reused sources. Requested guidance, realized question properties, medical validity, and downstream performance are distinct evidence levels. The training initialization and recipe are shared; selected steps appear in Table 3.

style refinement, and evidence-intensive reasoning practice. These are fixed allocations, not weights learned from Judge scores or an optimized mixture search.

D Corpus-Level Data Audit

Earlier corpus audits illustrate why item correctness and distributional fit require separate checks. The Case Report-derived corpus has 95.8% judged correctness and 99.6% unique answers, but overrepresents vignettes and underrepresents dentistry (Table 5). The Exam-derived corpus reduces some interface mismatch but still has task-distribution total variation 0.6589; treatment-principle questions account for 1.07% of generated items versus 14.68% in the reference distribution. These audits motivate profile alignment; they are not evaluations of the final training corpora or estimates of individual feature effects.

Table 5: Audit of the MedMCQA-targeted Case Report dataset versus the reference distribution used in that audit. High local validity coexists with interface and subject-coverage mismatch.

Audit measure	**Case dataset**	**Audit reference**
Judged answer correctness	95.8%	–
Unique-answer rate	99.6%	–
Median stem length (words)	76	11
Vignette form	99.3%	16.2%
Dentistry coverage	5.1%	31.5%
Negative-question form	7.4%	18.4%

E Feature Ablation Protocol and Additional Results

E.1 Inputs and Scope of the Intervention

Full retains the original clinical-decision corpus (6,539 Case Report examples) or evidence-intensive corpus (4,925 Case Report–Exam examples). Each removal reuses its setting's source pool, patient bindings or audited propositions, and precomputed Judge guidance, without new retrieval or source replacement. The six families are defined in Section 2.1; their exact fields and masking dependencies are in Appendix F. Only guidance is withheld: source facts and common medical-validity requirements remain available.

E.2 Construction and training

Removal corpora use Gemini-3.1-Pro-Preview for source-grounded generation, one refinement pass with retained guidance, and independent medical review. The reviewer receives the source and candidate with permuted options, but not the proposed answer or target profile, and checks correctness, grounding, unique answerability, and text-only English output. Generation uses source records rather than Full's questions; accepted examples contain questions and answer labels without teacher reasoning traces.

Clinical-decision removal corpora contain 6,466 decision-boundary, 6,493 evidence-structure, 6,493 reasoning-operation, and 6,508 answer-contrast examples. Evidence-intensive removal corpora contain 4,553, 4,586, 4,522, 4,514, 4,539, and 4,339 examples in the six-family order. Option counts remain fixed at four and ten, respectively. All conditions use their Full reference's starting model and training recipe; the comparisons concern the resulting corpus variants, whose sizes and selected training steps can differ.

E.3 Cross-task ablation results

All conditions use the evaluation protocol in Appendix C. Checkpoints are selected by the corresponding primary task; evidence-intensive removals are selected among steps 120, 240, and 360. Tables 6 and 7 extend the primary-task results in Table 3 with all four benchmark scores and the selected steps.

Table 6: Clinical-decision feature ablations across all four benchmarks. Each row is one checkpoint selected by MedQA. Δ MedQA is removal minus Full in percentage points; bold marks the highest score in each evaluation column.

Guidance	Step	MedXpert	MedQA	MedMCQA	HLE_verified	Δ MedQA
Qwen3-8B baseline	0	16.53	71.72	62.18	24.19	–
Full	420	**19.71**	**78.40**	62.16	25.12	–
−Knowledge target	360	18.41	72.58	62.25	24.65	−5.81
−Decision boundary	100	17.55	73.29	61.77	25.12	−5.11
−Evidence structure	240	17.43	75.18	**63.14**	26.05	−3.22
−Reasoning operation	180	16.04	73.53	62.20	**26.98**	−4.87
−Answer contrast	60	16.82	73.13	62.20	23.26	−5.26
−Assessment interface	20	16.24	73.06	62.73	25.58	−5.34

Table 7: Evidence-intensive feature ablations across all four benchmarks. Each row is one checkpoint selected by MedXpert. Δ MedXpert is removal minus this setting's Full in percentage points; bold marks the highest score in each evaluation column.

Guidance	Step	MedXpert	MedQA	MedMCQA	HLE_verified	Δ MedXpert
Qwen3-8B baseline	0	16.53	71.72	62.18	24.19	–
Full	360	**19.22**	71.33	61.03	22.79	–
−Knowledge target	240	18.20	72.90	**62.56**	27.44	−1.02
−Decision boundary	240	18.57	**73.37**	61.89	24.19	−0.65
−Evidence structure	360	17.80	70.54	60.51	24.65	−1.43
−Reasoning operation	360	17.96	71.88	61.18	27.44	−1.27
−Answer contrast	360	18.49	73.21	61.92	22.79	−0.73
−Assessment interface	240	18.24	73.06	62.37	**28.84**	−0.98

Full leads on the primary task in both settings, but not on every transfer evaluation. The paired analysis below applies to clinical-decision synthesis; evidence-intensive differences remain descriptive point estimates.

E.4 Paired uncertainty for feature-guidance contrasts

We compare the clinical-decision models on the same questions using 95% paired bootstrap intervals with 100,000 resamples (seed 20260923) and two-sided exact McNemar tests. Holm correction is

applied separately to six primary-task comparisons and 24 exploratory feature–task comparisons. These tests measure question-level uncertainty at the selected checkpoints, not variation across training seeds or checkpoint selection. They do not establish fixed-update causal effects or statistically resolve the ranking of guidance families.

Table 8: Paired MedQA contrasts on 1,273 questions. The difference is **Full minus removal**, as in Equation 4; positive values favor Full. n_{10} counts Full-correct/removal-wrong questions and n_{01} the reverse. p_{adj} is the exact McNemar p-value after Holm correction across these six comparisons.

Removed guidance	n_{10}	n_{01}	**Gain (pp)**	**95% paired CI**	p_{adj}
Knowledge target	152	78	+5.81	$[3.53, 8.09]$	7.27×10^{-6}
Decision boundary	148	83	+5.11	$[2.83, 7.46]$	1.03×10^{-4}
Evidence structure	111	70	+3.22	$[1.18, 5.34]$	0.00284
Reasoning operation	142	80	+4.87	$[2.59, 7.15]$	1.03×10^{-4}
Answer contrast	155	88	+5.26	$[2.91, 7.62]$	1.03×10^{-4}
Assessment interface	165	97	+5.34	$[2.91, 7.86]$	1.03×10^{-4}

Full gains 6.68 points over baseline on MedQA (95% paired CI: $[4.32, 9.03]$). All six Full–removal contrasts remain significant after correction (Table 8). Among removal–baseline comparisons, only evidence-structure removal retains a statistically supported gain after Holm correction across the seven trained-model–baseline comparisons: $+3.46$ points, CI $[1.34, 5.58]$, $p_{\mathrm{adj}} = 0.011$. The other five removal–baseline intervals include zero.

Table 9: Exploratory transfer contrasts: Full minus removal in percentage points, with 95% paired intervals. * denotes Holm-adjusted $p < 0.05$ across all 24 feature–task tests, including MedQA. Lack of a marker does not establish that a feature has no effect.

Removed guidance	**MedXpert**	**MedMCQA**	**HLE_verified**
Knowledge target	$+1.31$ $[-0.16, 2.78]$	-0.10 $[-1.29, 1.10]$	$+0.47$ $[-4.19, 5.12]$
Decision boundary	$+2.16$ $[0.57, 3.76]$	$+0.38$ $[-0.88, 1.65]$	0.00 $[-4.65, 4.65]$
Evidence structure	$+2.29$ $[0.78, 3.80]$	-0.98 $[-2.18, 0.22]$	-0.93 $[-5.58, 3.72]$
Reasoning operation	$+3.67$ $[2.20, 5.14]^*$	-0.05 $[-1.27, 1.17]$	-1.86 $[-6.98, 3.26]$
Answer contrast	$+2.90$ $[1.27, 4.53]^*$	-0.05 $[-1.34, 1.24]$	$+1.86$ $[-3.26, 6.98]$
Assessment interface	$+3.47$ $[1.84, 5.06]^*$	-0.57 $[-1.86, 0.72]$	-0.47 $[-5.12, 4.19]$

All MedMCQA and HLE_verified intervals in Table 9 include zero, so these transfer effects remain unresolved. A significant contrast on one task and a non-significant contrast on another does not itself establish a feature–task interaction.

F Detailed Labels for the Six Guidance Families

The tables specify the concrete guidance fields withheld in each ablation. Clinical-decision synthesis uses Case Reports; evidence-intensive synthesis uses Case Reports and Exam sources. Missing fields impose no additional requirement. These labels specify requested properties, not measured learning effects. Source facts, patient binding, medical-validity checks, and the four- or ten-option interface remain fixed across removals.

Dependencies and fields outside the six removals. Medical comparison axes are withheld when either knowledge-target or answer-contrast guidance is removed. The instruction to withhold an intermediate conclusion is removed with either evidence-structure or reasoning-operation guidance. Equivalent free-text instructions are masked consistently; source support, correctness, unique answerability, and route restrictions remain unchanged.

Table 10: **Knowledge target**: the medical content and learning objective the example should teach.

Compiled label	Definition and use
`target_disease_focus`; `medical_topic`	Disease, syndrome, or medical subject around which the question is organized. Case Report records use the former; evidence-intensive Exam records use the latter.
`target_task`; `assigned_task`	Requested clinical operation, such as diagnosis or test selection. Existing labels are reused; explicit exception or priority wording is separated from the topic-level operation.
`one_line_point`	One-sentence learning objective: the medical concept or relation to be taught. Used when present in the Case Report profile; it does not assert that a generic finding was observed in the patient.
`answer_semantic_type`	Kind of answer required, such as a disease, mechanism, test, drug, or management action. Exam records inherit the audited source-answer category.
`option_layer`	Medical category shared by the options, such as diagnoses or treatment choices. This defines their content domain, rather than how close the distractors should be.
`question_type`	Existing question-operation category supplied with the source/profile record, retained when available. It is distinct from narrative length or vignette style.
`commonness_tier`	Profile label for how typical the tested clinical knowledge is. Used in clinical-decision synthesis to retain its intended coverage.
`usmle_competency`	Licensing-style competency specified by the clinical-decision profile, such as diagnosis or management.
`question_operation`; `task_alias`	Topic-level operation recovered from the lead-in or reasoning label. The former is used in clinical-decision synthesis; the latter retains recognized operation labels in evidence-intensive synthesis.

Table 11: **Decision boundary**: when a clinical rule or action applies. Source-documented contraindications remain available.

Compiled label	Definition and use
`decision_scope`	Whether the answer identifies a fact or selects an action for a particular patient. Clinical-decision profiles distinguish standard and exception decisions. In the evidence-intensive compiler, the value `knowledge_fact` belongs to knowledge targeting; patient-decision scope belongs here.
`standard_vs_exception`	Whether the requested framing follows a standard rule, a source-supported exception, or a mixture. Exam synthesis defaults to a standard rule unless the source requires an exception.
`canonicality_scope`	Preference for a standard clinical rule rather than an idiosyncratic reported choice. The clinical-decision compiler sets this to `standard_clinical_rule`.
`priority_level`	Whether the requested action is initial, primary, next, or unspecified. Clinical-decision synthesis extracts this from the lead-in and learning objective.
`exception_focus`	Whether the requested decision specifically concerns a contraindication or alternative treatment. Used for evidence-intensive Case Report synthesis without inventing a new exception.

Table 12: **Evidence structure**: which source-supported clues are presented and how. The source itself is not shortened.

Compiled label	Definition and use
`evidence_slot_sequence`	Ordered, content-independent clue types, such as symptoms, time course, examination, and laboratory findings. Normalized from stored evidence slots or their source-specific adaptations. Only available source-supported clues may fill a slot; Exam records may have no prescribed sequence.
`decisive_clue_count`	Desired number of answer-relevant clues. Clinical-decision guidance requests 2–4; this is a construction preference, not an observed count.
`neutral_detail_count`	Desired number of non-decisive contextual details. Clinical-decision guidance requests 0–2 to limit unnecessary narrative content.
`information_rule`	Instruction to use the minimum sufficient answer-relevant source evidence. Used in evidence-intensive synthesis instead of a fixed clue-count band.
`intermediate_conclusion_disclosure`	Whether the stem should leave an intermediate explanation for the learner to infer. The clinical-decision instruction says not to state that explanation directly. It is shared with reasoning operation and removed if either family is withheld.

Table 13: **Reasoning operation**: how the visible evidence should lead to an answer. These are item-design requirements, not teacher reasoning traces or model inference-mode settings.

Compiled label	Definition and use
`cognitive_operations`	Content-independent operations derived from the stored reasoning mode: for example, recognize a pattern then apply a rule, integrate clues then distinguish competing interpretations, or calculate then compare with a criterion.
`reasoning_depth`	Requested number or dependency of inference steps. Clinical-decision synthesis requests at least two linked steps. Evidence-intensive Case Reports retain recoverable one-, two-, or multi-step requirements, otherwise unspecified; Exam synthesis requests two linked steps only when source-supported.
`action_order`	Whether the reasoning must resolve temporal priority among actions. The evidence-intensive Case Report compiler sets this for emergency/sequencing tasks; otherwise it is unspecified.
Shared conclusion instruction	The clinical-decision `intermediate_conclusion_disclosure` field in Table 12 is removed with either evidence structure or reasoning operation, preventing that dependency from restoring omitted guidance.

Table 14: **Answer contrast**: the distinctions between the correct answer and plausible alternatives. Option count and unique-answer requirements remain fixed.

Compiled label	Definition and use
`option_relation`	How alternatives relate to one another: semantically comparable choices in clinical-decision synthesis, and parallel, plausible near-neighbor alternatives in evidence-intensive synthesis.
`comparison_dimensions`	Number of attributes along which the options should be comparable, derived from the stored option axes with a minimum of one. It does not by itself reveal the medical content of those axes.
`distractor_relations`	Generic near-miss design: alternatives explain some but not all findings, or provide coherent competing interpretations rather than unrelated choices. Explicitly compiled for clinical-decision synthesis.
`medical_comparison_axes`	Named medical dimensions, such as mechanism, anatomical site, or treatment timing, taken from stored option axes. This shared instruction depends on both knowledge target and answer contrast, and is withheld when either is removed. Used when those axes are available.

Table 15: **Assessment interface**: how the question is presented, without changing medical facts.

Compiled label	Definition and use
`vignette_mode`	Whether the item is presented as a patient vignette. Clinical-decision guidance requests a clinical vignette.
`stem_length_words`	Preferred stem-length range. Clinical-decision guidance requests 70–220 words; this is not a change to the model's generation-token budget.
`clinical_setting`	Requested care setting from the clinical-decision profile. It may be used only if compatible with the actual source patient; it cannot justify inventing an encounter.
`style`	Narrative and lead-in convention. Clinical-decision guidance requests a licensing-style vignette with a final question.
`presentation_style`	Evidence-intensive presentation preference: a clinical vignette or direct question for Case Reports, and a concise board-style item for Exam sources.

G Supplementary Analysis of Guidance Utility

G.1 Question Matching and Complete Clinical-Operation Counts

The analysis in Section 3.7 reuses stored predictions for CD Full and knowledge-target removal, and EI Full and evidence-structure and reasoning-operation removal. These are the largest aggregate primary-task contrasts in Table 3. No responses are regenerated. Matching question text and reference answers joins all 1,273 MedQA and 2,450 MedXpertQA questions to their existing clinical-operation annotations. These are coarse, previously generated labels, not clinician-adjudicated error explanations. We compare the same questions within each contrast; the category composition differs between benchmarks.

Table 16: **Complete clinical-operation counts.** Each cell reports a correct-answer count, except N, the number of questions. KT, ES, and RO denote removal of knowledge-target, evidence-structure, and reasoning-operation guidance. A dash indicates that the category has no questions. The order matches the combined category sizes used in the main-text summary.

	CD / MedQA			EI / MedXpert			
Clinical operation	N	Full	−KT	N	Full	−ES	−RO
Etiology and mechanism	295	228	207	311	53	56	52
Diagnosis and differentiation	205	173	173	263	41	36	40
Non-disease topics	156	120	111	290	50	60	51
Treatment principles	132	115	98	255	54	53	56
Next-investigation selection	49	38	37	238	57	51	51
Drug selection	72	56	51	201	49	36	36
Test interpretation and assessment	75	60	50	152	25	22	25
Structure and function	95	71	66	130	17	11	13
Complications and associations	46	29	27	93	6	6	9
Clinical manifestations	31	21	23	97	11	7	16
Emergency priorities and sequencing	33	20	18	93	23	24	21
Surgery and procedure selection	22	20	19	92	20	16	16
Contraindications and alternatives	16	11	11	86	9	10	10
Prevention and prognosis	18	18	14	57	21	15	19
Non-drug intervention	14	10	8	37	15	17	12
Care setting and disposition	10	6	8	36	14	11	9
Concepts and classification	4	2	3	16	5	4	3
Complex decision trade-offs	–	–	–	3	1	1	1
Total	1,273	998	924	2,450	471	436	440

Pairing also exposes changes hidden by total scores. For the CD knowledge-target contrast, Full alone answers 152 questions correctly and removal alone answers 78, yielding a net difference of 74. The corresponding counts for EI evidence-structure removal are 199 versus 164, and for reasoning-operation removal 191 versus 160. Full therefore does not solve a strict superset of the removal model's questions. Equal diagnosis totals in CD conceal 12 changed answers in each direction. These are benchmark scoring outcomes, not a new expert assessment of the medical reasoning in each response.

G.2 Archived Construction Examples

Figure 7 summarizes two archived initial–refined pairs. They were selected to explain the construction requirements, not to estimate a refinement success rate. They are historical revisions involving multiple checks, not single-feature interventions.

CD: choosing a supported learning target. The report *Isolated Right-sided Pulsus Alternans in Pulmonary Arterial Hypertension* (DOI: `10.2169/internalmedicine.50.5285`) describes a 79-year-old woman with systemic sclerosis, pulmonary artery pressures of 99 mmHg systolic and 51 mmHg mean, a non-elevated wedge pressure, and isolated right-sided pulse alternation. The source audit excludes questions about the pulse alternation's mechanism, which the report leaves unresolved. The initial item nevertheless asks which hemodynamic alteration explains the abnormal tracings.

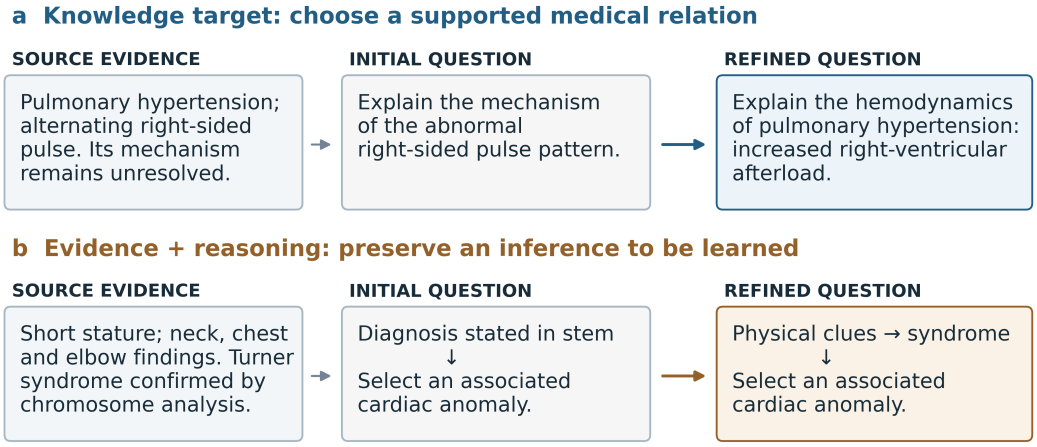

Figure 7: **What guidance changes during construction.** Schematic summaries of archived revisions. (a) CD redirects an unsupported explanation of a rare pulse pattern to a supported hemodynamic learning target. (b) EI removes an explicit diagnosis while retaining the physical clues, restoring an inference before an association question. These illustrate source-constrained construction decisions, not the isolated downstream effect of a feature.

The revision instead asks which alteration characterizes the established pulmonary hypertension and restores the documented age and pressure values. This changes the learning objective within the evidence boundary, rather than merely correcting the answer.

EI: retaining clues while restoring an inference. *PCR–PRINS–FISH Analysis of Structurally Abnormal Sex Chromosomes in Eight Patients with Turner Phenotype* (*Clinical Genetics*, 2001, 60:385–392) supplies one bound patient's short stature, webbed neck, chest-wall and elbow findings, chromosomal diagnosis, and unspecified congenital cardiopathy. The initial item states the diagnosis before asking about an associated cardiac anomaly. The revision removes the explicit karyotype/diagnosis but retains the physical clues, requiring syndrome recognition before association selection. An undocumented auscultatory clue is disallowed rather than invented. The anomaly is queried as expected clinical knowledge, not asserted as an observed finding in that patient; the answer remains unchanged.

Scope of the second-backbone comparison. The Qwen3-30B-A3B gains in Table 2 correspond to 48 additional correct MedXpert answers, 48 MedQA answers, 52 MedMCQA answers, and three HLE_verified answers. Each row is a separate trained model's designated evaluation, not one model's joint score vector. Original-model evaluation uses tensor parallelism of eight and trained-checkpoint evaluation uses four. This extends the corpus-level results within Qwen3, but neither replicates the guidance ablations nor establishes their ranking on A3B.