MEDGUIDEX: INTERNALIZING DECISION LOGIC FROM EXECUTABLE GUIDELINES INTO LARGE LANGUAGE MODELS FOR CLINICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical practice guidelines (CPGs) encode evidence-based decision logic that clinicians apply by evaluating patient variables, conditional criteria, and recommendation rules. However, existing methods often use CPGs as free-text training data or retrieval sources, underutilizing their procedural decision structure. To better exploit this structure, we introduce a guideline-derived training pipeline that transforms CPG recommendations into executable clinical decision logic and uses it to generate factual and counterfactual question-answering data. These data teach models both guideline-supported decisions and how decisions change under different patient conditions. Post-training a medical LLM on the generated data yields *MedGuideX*. Across four clinical reasoning benchmarks, *MedGuideX* achieves a 10.42% relative improvement in average accuracy. Physician evaluation further shows that *MedGuideX* better recovers clinician-authored reasoning steps and produces physician-preferred rationales in faithfulness, validity, completeness, and clarity. Overall, our results show that executable decision logic from CPGs can be transformed into scalable supervision for building reliable medical LLMs.

1 INTRODUCTION

Large language models (LLMs) (Singh et al., 2025; Yang et al., 2025) have shown strong potential in medical domain, including electronic health records understanding, clinical case reasoning, and medical decision support (Cao et al., 2026; Wu et al., 2025b; Lai et al., 2025). However, reliable clinical reasoning remains challenging. It requires models to integrate heterogeneous patient evidence, apply domain knowledge, compare plausible clinical decisions, handle uncertainty, and follow evidence-based decision logic (Bowen, 2006; Nendaz & Perrier, 2012; Sox et al., 2024).

Clinical practice guidelines (CPGs) offer a natural source of such decision logic. In clinical practice, clinicians apply guidelines by identifying patient variables, evaluating conditional criteria, and following recommendation rules. Thus, beyond textual medical knowledge, CPGs encode procedural decision structures for diagnosis, treatment, and disease management.

To better exploit this structure, we propose a guideline-grounded training pipeline for building LLMs with stronger clinical reasoning ability. We first collect high-quality, publicly available CPGs and transform their recommendations into executable functions that represent structured clinical decision logic. Each function operates over patient variables and produces guideline-consistent decisions, enabling controlled data generation and automatic verification. Based on these functions, we generate factual and counterfactual question-answering instances. Factual instances teach models guideline-supported decisions, while counterfactual instances teach how decisions should change when key patient conditions are modified.

Using this pipeline, we train *MedGuideX*, a medical LLM designed to internalize guideline-grounded clinical decision logic. Specifically, we post-train the base model with supervised fine-tuning (SFT) and reinforcement learning (RL) on the generated factual and counterfactual data. Experiments on four clinical reasoning benchmarks show that *MedGuideX* substantially improves over its base model and achieves strong performance among open-source medical LLMs. Compared with Qwen3.5-9B (Yang et al., 2025), *MedGuideX*-9B achieves relative improvements of 26.64%, 9.45%, 4.41%, and 11.67% on MedCaseReasoning (Wu et al., 2025b), MedQA (Jin et al., 2021), MIMIC-CDM-FI (Hager et al.,

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.

2024), and ER-Reason (Mehandru et al., 2025), respectively. Notably, the larger relative gains appear on the lower-accuracy benchmarks, MedCaseReasoning and ER-Reason, suggesting that guideline-derived supervision is particularly useful for challenging reasoning settings. Physician evaluation further shows that *MedGuideX* better recovers clinician-authored reasoning steps and produces physician-preferred rationales in faithfulness, validity, completeness, and overall quality.

In summary, our contributions are:

- We propose a guideline-derived post-training pipeline that transforms CPGs into executable clinical decision logic and uses it to generate factual and counterfactual QA supervision.
- We train *MedGuideX*, a medical LLM that internalizes guideline-grounded clinical decision logic through SFT and RL.
- We conduct experiments on four medical reasoning benchmarks, showing that *MedGuideX* improves over the base model and similarly sized medical LLMs, while producing higher-quality clinical rationales.

2 Related Work

2.1 LLMs for Clinical Reasoning

Medical LLMs are trained on large-scale medical corpora, clinical notes, and case reports (Chen et al., 2023; Han et al., 2023; Labrak et al., 2024; Garcia-Gasulla et al., 2025; Wu et al., 2025b), and enhanced through retrieval, multi-agent diagnosis, or clinical-feedback post-training (Cao et al., 2026; Li et al., 2025; Chen et al., 2024; Lai et al., 2025). These heterogeneous sources broaden medical knowledge and case-level reasoning, but often provide implicit, noisy, or incomplete supervision weakly aligned with explicit clinical decision procedures (Gu et al., 2025; Li et al., 2026; Yoo & Woo, 2025; Yang et al., 2024).

2.2 Clinical Practice Guidelines for LLMs

CPG-based methods retrieve guideline passages for prompting (Schubert et al., 2025; Deng et al., 2026; Oniani et al., 2024; Li et al., 2023a), construct guideline-grounded SFT rationales or train directly on guideline text (Staniek et al., 2025; Chen et al., 2023), and use guideline constraints as RL rewards (Tziakouri & Menolascina, 2025; Gu et al., 2025). These uses expose guideline content but do not explicitly operationalize its variables, conditions, and decision rules, leaving the internal logic underused as scalable supervision. In contrast, our framework compiles this logic into executable functions that generate factual and counterfactual QA supervision.

2.3 Post-Training and Executable Counterfactual Reasoning

Intermediate reasoning steps such as chain-of-thought can elicit complex reasoning (Wei et al., 2022a; Suzgun et al., 2023; Wei et al., 2022b; Cao, 2024), while post-training seeks to internalize such behavior. SFT uses filtered self-generated solutions (Yuan et al., 2023; Tong et al., 2024) or rationales distilled from stronger models (Muennighoff et al., 2025), whereas RL (Kaelbling et al., 1996) optimizes task-level objectives beyond next-token prediction (Lightman et al., 2024; Uesato et al., 2022; Guo et al., 2025). Large-scale RL and its extensions to search and structured data further demonstrate this paradigm (Guo et al., 2025; Jiang et al., 2025a; Jin et al., 2025; Cao et al., 2025). Counterfactual reasoning can also strengthen model reasoning and reveal failures missed by standard QA (Chen et al., 2025; You et al., 2026; Vashishtha et al., 2025). Our work combines these directions by deriving verifiable factual and counterfactual clinical supervision from executable, evidence-based guidelines.

3 Methodology

3.1 Logic Formulation

We define a guideline recommendation as an actionable clinical decision rule extracted from a complete CPG. A recommendation specifies how a clinical decision should be made under certain

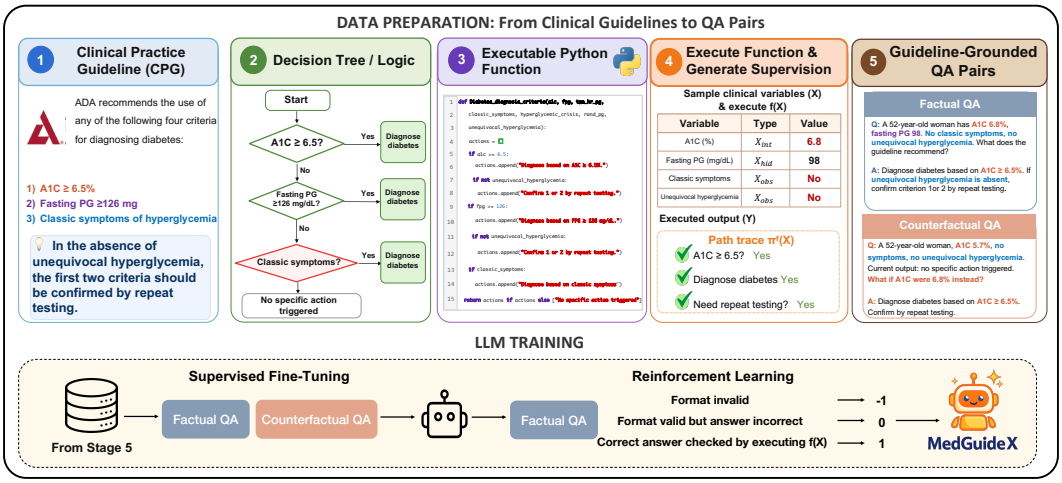

Figure 1: Overview of *MedGuideX*. **Top**: We transform raw CPGs into executable Python functions f via an intermediate decision tree, then sample clinical variables X and execute $f(X)$ to deterministically label both factual and counterfactual QA instances. **Bottom**: We post-train an LLM with SFT on the mixed QA and RL on factual QA, where re-executing f provides the reward. This design turns each guideline into an executable verifier that grounds both labels and rewards in its decision logic.

patient conditions, such as diagnosis, treatment, screening, or disease management. For example, a recommendation may state that a patient with severe symptoms should be referred for further evaluation.

We formalize each guideline recommendation as a function

$$Y = f(X), \qquad f : \mathcal{X} \to \mathcal{Y}, \tag{1}$$

where $X = (x_1, \ldots, x_n) \in \mathcal{X}$ denotes a vector of clinical variables, f denotes the conditional decision logic encoded by the recommendation, and $Y \in \mathcal{Y}$ is the guideline-prescribed output. We instantiate f as a finite decision tree T_f, whose internal nodes are atomic predicates $a \in \mathcal{A}_f$ over X and whose leaves are outputs in $\mathcal{Y}$. Each predicate takes the form $a_i = \mathbf{1}[g_i(X)]$, where g_i is a guideline-defined condition, such as *age* ≥ 65 or *eGFR* < 30.

Given an input X, executing T_f activates a path

$$\pi_f(X) = (a_{i_1}, a_{i_2}, \ldots, a_{i_k}) \subseteq \mathcal{A}_f, \tag{2}$$

whose predicates jointly select the leaf output $Y = f(X)$. This path is the verifiable unit of guideline logic that we aim for the model to internalize: it should not only predict the correct output, but also reason through a path consistent with $\pi_f(X)$.

We next describe how we collect CPGs and construct factual and counterfactual QA data from them (§ 3.2), followed by how we train *MedGuideX* using SFT and RL (§ 3.3). Figure 1 illustrates the full pipeline.

3.2 Data Construction

The central artifact of data preparation is an executable implementation of f, whose control flow mirrors the decision tree T_f. This executable form enables downstream supervision by deterministically labeling synthesized questions through direct execution and verifying model predictions during training by re-executing f on the model-stated intermediate variables. In this formulation, the inputs correspond to structured patient information and clinical scenarios, while the executable function f represents structured clinical reasoning grounded in guideline decision logic.

Guideline Curation

Our initial guideline source is an open CPG collection based on the corpus used to train MEDITRON (Chen et al., 2023). However, many documents in this collection are noisy, low quality, or

near-duplicate guidelines, which can substantially degrade the quality of downstream QA data. We therefore apply a curation pipeline to retain high-quality guidelines.

We first restrict the corpus to guidelines from U.S.-based sources, since clinical recommendations may vary across countries, healthcare systems, and organizations. Specifically, we retain guidelines from the Centers for Disease Control and Prevention (CDC)[1] and PubMed[2]. We then use an LLM to extract structured metadata for each guideline, including the disease or drug, target age group, race, gender, and publication date. Guidelines with identical metadata are treated as duplicates, and only the most recent version is retained. In addition, we instruct the LLM to discard incomplete documents directly. After curation, we obtain a filtered subset of CPGs.

EXECUTABLE TRANSFORMATION

Recommendation Extraction. We first split each document into recommendation-oriented chunks, where each chunk contains one or more complete guideline recommendations. This produces a set of guideline recommendation passages. An LLM extractor then identifies recommendation candidates from each chunk. For each candidate, we ask the extractor to identify the target population, clinical condition, recommended action, relevant exceptions, and evidence grade, when available. We then validate these candidates and retain only usable recommendations. Specifically, we discard candidates that do not describe a concrete clinical action or cannot be expressed as a condition-action rule. We also remove near-duplicates, where multiple recommendations describe highly similar populations, conditions, and actions.

Decision-Tree Validation. Each retained recommendation is converted into a decision tree T_f, which specifies the required input variables, decision conditions, and final outputs. An LLM validator then checks whether the tree is complete, whether each condition is clear, whether every branch leads to an output, and whether all variables are supported by the source guideline.

Compilation to Executable Function. Each validated tree is compiled into an executable Python function that takes the variables X as input and returns the guideline output $f(X)$. We further check whether the function is syntactically correct, executable on sampled inputs, and consistent with the original decision tree.

QA SYNTHESIS

Factual QA Synthesis. For each executable function f, we sample complete clinical variable assignments X and execute f to obtain the guideline output $Y_{\text{obs}} = f(X)$. A naive sampling strategy would produce too many easy or default cases, such as no-action recommendations. To avoid this imbalance, we enforce two constraints: (1) ***Path coverage***: the generated data should cover all decision conditions in the tree, and (2) ***Output balance***: no-diagnosis outputs, meaning that no diagnosis is recommended for the current inputs, should not dominate the dataset.

After applying the coverage and balancing constraints, we obtain the factual QA set. For each sample, we use an LLM to generate a step-by-step reasoning trace from the underlying Python function f, input variables X, and executed output Y_{obs}. The reasoning trace verbalizes the executed decision path $\pi_f(X)$ and is stored alongside the QA pair for SFT training.

Counterfactual QA Synthesis. We further generate counterfactual QA data to train the model to reason about hypothetical changes in patient conditions. For each counterfactual example, we first sample a complete variable assignment X over the inputs of f and execute f to obtain the factual outcome $Y_{\text{obs}} = f(X)$. We then partition the variables into three disjoint parts:

- X_{obs}: observed variables that are shown and remain unchanged.
- X_{hid}: hidden variables that are not shown.
- x_{int}: a single observed variable modified by the intervention, meaning that its value is changed while all other observed variables are held fixed.

[1] https://www.cdc.gov/
[2] https://pubmed.ncbi.nlm.nih.gov/

We write $\hat{x}_{\text{int}}$ for the value of x_{int} after intervention. Only x_{int} changes, while X_{obs} and X_{hid} are held fixed. The factual and counterfactual outcomes can be written as

$$\begin{aligned} Y_{\text{obs}} &= f\big(X_{\text{obs}}, X_{\text{hid}}, x_{\text{int}}\big), \\ \hat{Y}_{\text{cf}} &= f\big(X_{\text{obs}}, X_{\text{hid}}, \hat{x}_{\text{int}}\big). \end{aligned} \tag{3}$$

By construction, Y_{obs} and $\hat{Y}_{\text{cf}}$ differ only through the intervention on x_{int}.

The model receives X_{obs}, the identity of the hidden variables, the original and intervened values x_{int} and $\hat{x}_{\text{int}}$, and the factual outcome Y_{obs}. It must then predict $\hat{Y}_{\text{cf}}$ through three steps:

1. **Abduction.** Infer hidden values X_{hid} that are consistent with the observed context and reproduce Y_{obs} when executed through f.
2. **Intervention.** Replace x_{int} with $\hat{x}_{\text{int}}$ while keeping X_{obs} and X_{hid} fixed.
3. **Prediction.** Execute f on the resulting complete input to obtain the counterfactual outcome $\hat{Y}_{\text{cf}}$.

At the data level, we execute f on the factual input, apply the intervention, and execute f again to generate interventional scenarios. We retain only cases where the intervention changes the outcome and discard those with unchanged outputs. Each counterfactual sample is paired with a reasoning trace generated by an LLM, which verbalizes the three steps of abduction, intervention, and prediction over the executable function f. Overall data pipeline statistics are shown in Appendix C.

3.3 Post-Training

We post-train *MedGuideX* using prepared factual and counterfactual QA sets.

Optimization objectives. For each guideline recommendation, we obtain an executable function f. Given a complete patient scenario X, the function returns the guideline output

$$Y = f(X). \tag{4}$$

A factual QA instance asks the model to determine the guideline output for the complete patient scenario X. For factual examples, each oracle trajectory contains the patient scenario, the executed guideline output, and a rationale that explains why the scenario leads to that output Y.

For counterfactual QA examples, each trajectory follows an abduction–intervention–prediction structure. The prompt providesX_{obs}, the original value x_{int}, the intervened value $\hat{x}_{\text{int}}$, and the factual guideline output Y_{obs}; the hidden variable is not shown to the model. The model must first infer a hidden value that makes the factual observation consistent with the executable guideline, then apply the intervention while keeping the inferred hidden value fixed, and finally predict the counterfactual guideline output $\hat{Y}_{\text{cf}}$.

The optimization objective is to train the model so that its predicted guideline output $\hat{Y}$ matches the oracle output Y as closely as possible in both factual and counterfactual QA, thereby improving its ability to perform clinical reasoning aligned with the CPGs from which the synthetic data are derived.

SFT. The SFT stage trains the model to imitate oracle trajectories synthesized from the executable guideline functions. This stage serves as a cold start for subsequent RL training by injecting guideline-grounded clinical knowledge and exposing the model to a structured clinical reasoning framework.

RL. After SFT, we apply GRPO (Shao et al., 2024) to optimize the model on its own sampled responses. For a sampled response o, we define a format reward $r_{\text{fmt}}(o)$ and an answer reward $r_{\text{answer}}(o)$. The format reward checks whether o satisfies the required response format, while the answer reward checks whether the parsed answer is correct. The final reward is

$$r(o) = \begin{cases} -1, & r_{\text{fmt}}(o) = -1, \\ r_{\text{answer}}(o), & r_{\text{fmt}}(o) = 0. \end{cases} \tag{5}$$

Here, $r_{\text{fmt}}(o) = -1$ indicates an invalid response format, $r_{\text{fmt}}(o) = 0$ indicates a valid format, and $r_{\text{answer}}(o) \in \{0, 1\}$ indicates whether the task-specific correctness check passes.

For factual prompts, correctness depends only on the final guideline output. Let $\hat{Y}(o)$ denote the final answer parsed from a sampled response o. We define the factual answer reward as

$$r_{\mathrm{F}}(o) = \mathbb{1}\left[\hat{Y}(o) \equiv Y^{\star}\right]. \tag{6}$$

For counterfactual prompts, correctness requires more than the final answer. Let $\hat{X}_{\mathrm{hid}}(o)$ denote the hidden value inferred by the model, and let $\hat{Y}_{\mathrm{cf}}(o)$ denote its final counterfactual answer. The response is counted as correct only if the inferred hidden value matches the intended hidden state, the inferred hidden value together with the original factual context reproduces Y_{obs}, and the final counterfactual answer matches the executed counterfactual label:

$$\begin{aligned} r_{\mathrm{CF}}(o) = \mathbb{1}&\left[\hat{X}_{\mathrm{hid}}(o) = X^{\star}_{\mathrm{hid}}\right] \\ &\cdot \mathbb{1}\left[f\left(X_{\mathrm{obs}}, \hat{X}_{\mathrm{hid}}(o), x_{\mathrm{int}}\right) = Y_{\mathrm{obs}}\right] \\ &\cdot \mathbb{1}\left[\hat{Y}_{\mathrm{cf}}(o) = Y^{\star}_{\mathrm{cf}}\right]. \end{aligned} \tag{7}$$

Thus, the counterfactual reward requires the model to recover the intended hidden value and predict the correct counterfactual output.

Training procedure. We perform SFT on the balanced factual and counterfactual QA mixture, then initialize GRPO from the SFT checkpoint. The reward implementation supports both factual and counterfactual QA. Our final *MedGuideX* configuration uses mixed factual and counterfactual SFT followed by factual GRPO, as selected by the ablation study in Appendix F. Counterfactual GRPO is evaluated as an alternative configuration in the same ablation.

4 Experiments

4.1 Experimental Setup

We evaluate *MedGuideX* on four clinical reasoning benchmarks spanning medical exam questions and real-world case-based diagnostic reasoning: MedQA (Jin et al., 2021), MedCaseReasoning (Wu et al., 2025b), MIMIC-CDM-FI (Hager et al., 2024), and ER-Reason (Mehandru et al., 2025). These four benchmarks differ in data source, task format, reasoning depth, and difficulty, enabling a more comprehensive evaluation of *MedGuideX*. Details and examples of these benchmarks are provided in Appendix D. Notably, none of the training sets from these benchmarks are used during the training of *MedGuideX*.

We compare *MedGuideX* with three groups of baselines, with sources listed in Table 1. *Frontier LLMs* include strong proprietary or large-scale general-purpose models and serve as an upper-performance reference. To comply with institutional data governance requirements, all proprietary model inference is conducted through Azure AI. *Open-source Medical LLMs* are models specialized for the medical domain through medical pretraining or fine-tuning, and constitute the most directly comparable baselines to *MedGuideX* at a similar scale. *Base models* are the untuned Qwen3.5-4B and Qwen3.5-9B backbones used by *MedGuideX*. We also compare against methods that leverage CPGs in different ways (Appendix H). The main results report final-answer accuracy, measured by exact match or LLM-as-a-judge evaluation. Additional experimental details are provided in Appendix E.

4.2 Main Results

Table 1 reports accuracy on four clinical reasoning benchmarks. Overall, *MedGuideX* consistently improves performance at both model scales. At the 9B scale, *MedGuideX* improves over Qwen3.5-9B by 26.64%, 9.45%, 4.41%, and 11.67% on MedCaseReasoning, MedQA, MIMIC-CDM-FI, and ER-Reason, respectively, increasing the average score by 10.42%. The 4B model shows the same trend, with relative improvements of 23.35%, 6.39%, 2.49%, and 9.75%, and an average-score improvement of 7.63%. These results show that guideline-derived post-training provides a consistent improvement signal across both model scales.

The relative gains are largest on MedCaseReasoning and ER-Reason, the two benchmarks where the backbone accuracy is lowest. This suggests that guideline-derived decision supervision is particularly

Table 1: Main results on four clinical reasoning benchmarks (accuracy, %). **Bold** marks the best score in each column, and *italics* indicate the second-best score among open-source medical LLMs. Relative gains over the corresponding base model are shown in green. *MedGuideX* achieves strong performance across benchmarks.

Model	MedCase Reasoning	MedQA	MIMIC-CDM-FI	ER-Reason	Average
Frontier LLM					
GPT-5.0 (Singh et al., 2025)	54.96	87.43	85.10	33.70	65.30
DeepSeek-V4 (Liu et al., 2024)	52.51	88.92	90.10	24.00	63.88
Kimi-2.6 (Team et al., 2025)	23.63	79.26	74.60	34.40	52.97
Claude-Haiku-4.5 (Anthropic, 2025)	40.02	48.08	81.60	34.30	51.00
Open-source Medical LLM					
Clinical-R1-3B (Gu et al., 2025)	13.60	52.55	62.30	19.10	36.89
MediPhi-Instruct-4B (Corbeil et al., 2025)	17.06	53.34	72.60	22.20	41.30
MedGemma-1.5-4B (Sellergren et al., 2025)	16.39	67.79	80.10	20.80	46.27
Hulu-Med-7B (Jiang et al., 2025b)	31.22	72.98	78.70	22.20	51.28
Lingshu-7B (Xu et al., 2025)	23.86	64.49	78.00	26.20	48.14
Llava-Med-v1.5-7B (Li et al., 2023b)	16.72	44.78	60.10	21.20	35.70
MedAlpaca-7B (Han et al., 2023)	12.26	38.81	41.00	11.00	25.77
BioMistral-7B (Labrak et al., 2024)	17.28	45.80	35.60	24.80	30.87
HuatuoGPT-o1-8B (Chen et al., 2024)	19.29	71.09	68.10	22.00	45.12
Llama-Aloe-Beta-8B (Garcia-Gasulla et al., 2025)	21.18	64.65	80.80	**29.10**	48.93
MedReason-8B (Wu et al., 2025a)	27.76	70.86	73.00	22.40	48.51
Base LLM & Standard Baselines					
Qwen3.5-4B (Yang et al., 2025)	28.65	72.43	80.40	23.60	51.27
+ RAG with Guidelines	28.54	73.45	81.90	26.30	52.55
+ 3-Shot In-Context Learning	28.65	71.41	81.40	26.10	51.89
Qwen3.5-9B (Yang et al., 2025)	32.66	73.21	81.60	25.70	53.29
+ RAG with Guidelines	31.88	74.71	82.50	27.40	54.12
+ 3-Shot In-Context Learning	33.11	74.94	80.10	28.20	54.08
+ CPGPrompt (Deng et al., 2026)	14.16	59.15	67.60	19.90	40.20
+ Fine-tuning with CPG	28.87	65.04	78.40	21.50	48.45
+ RL with CPG-Derived Process Rewards	37.68	75.73	83.40	27.80	56.15
Ours					
***MedGuideX*-4B**	*35.34*$_{+23.35\%}$	*77.06*$_{+6.39\%}$	*82.40*$_{+2.49\%}$	25.90$_{+9.75\%}$	*55.18*$_{+7.63\%}$
***MedGuideX*-9B**	**41.36**$_{+26.64\%}$	**80.13**$_{+9.45\%}$	**85.20**$_{+4.41\%}$	*28.70*$_{+11.67\%}$	**58.85**$_{+10.42\%}$

useful in more challenging reasoning settings. At the same time, the gains on MIMIC-CDM-FI are smaller, likely because this benchmark already has high backbone performance and requires additional abilities beyond guideline decision logic, such as synthesizing noisy EHR evidence and mapping clinical findings to benchmark-specific labels. Thus, *MedGuideX* improves broadly, but the magnitude of improvement depends on how directly each benchmark matches the decision-logic supervision used in training.

Compared with open-source medical LLMs, *MedGuideX* achieves strong performance. *MedGuideX*-9B obtains the best average score among all open-source medical LLMs in the table and is the strongest open-source model on MedCaseReasoning, MedQA, and MIMIC-CDM-FI, while remaining competitive on ER-Reason. *MedGuideX*-4B also exceeds the average performance of all listed open-source medical LLMs, showing that the proposed training pipeline is effective even at a smaller scale. Although *MedGuideX*-9B remains below the strongest proprietary systems on average, it narrows the gap substantially while using a compact open model.

The guideline-based baselines further clarify the importance of executable supervision. Retrieval-augmented guideline prompting and in-context guideline demonstrations provide only modest improvements over Qwen3.5-9B, increasing the average score by 1.56% and 1.48%, respectively. These methods expose the model to guideline text or guideline-style examples at inference time, but do not train the model to internalize the underlying decision logic. CPGPrompt (Deng et al., 2026), which constructs a decision-tree prompt from retrieved guideline recommendations, performs substantially worse in our setting, decreasing the average score by 24.56%. Direct fine-tuning on raw CPG text also leads to negative transfer, decreasing the average score by 9.08%, suggesting that textual exposure to guidelines alone is insufficient for learning operational clinical decision rules.

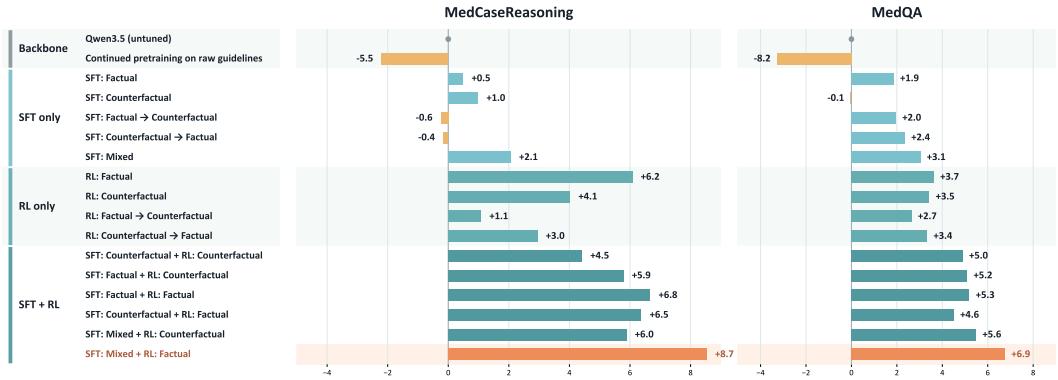

Figure 2: Training strategy ablation on data composition and training phases. Each group of bars corresponds to a configuration specifying the data used in the SFT and RL phases. *Mixed* denotes a jointly trained corpus with equal factual and counterfactual proportions, and $A \rightarrow B$ denotes applying data type A before data type B within a single phase. The best overall configuration uses mixed-data SFT followed by factual RL.

The strongest guideline-based baseline is RL with CPG-derived process rewards, which improves the average score by 5.37% over Qwen3.5-9B. However, it still underperforms *MedGuideX*, whose average relative improvement is 10.42%. This baseline uses an LLM judge to assess consistency with retrieved guideline recommendations, but it does not execute guideline functions or verify counterfactual decision behavior. The gap between this baseline and *MedGuideX* suggests that executable guideline functions provide a more precise and effective supervision signal than retrieval-based or judge-only uses of CPGs. A controlled qualitative case study of *MedGuideX* with GPT-5.0 and MedReason-8B is presented in Appendix I.

Ablation study Detailed ablations are reported in Appendix F. The best configuration applies mixed factual/counterfactual SFT followed by factual RL, yielding gains of 8.7 points on MedCaseReasoning and 6.9 points on MedQA over Qwen3.5-9B. This configuration is used for *MedGuideX*-9B, supporting our design choice of converting guideline decision logic into executable QA supervision and training strategies.

4.3 Paired Answer Transition Analysis

To better understand the source of the performance gains, we conduct a paired comparison between *MedGuideX* and its Qwen3.5 backbone on the same test instances. For each benchmark, we categorize answer transitions into three groups: *Corrected*, *Regressed*, *Unchanged*.

Table 2: Paired answer transition analysis between Qwen3.5 and *MedGuideX*. Corrected denotes cases where *MedGuideX* changes a wrong Qwen3.5 answer into a correct one, while Regressed denotes the opposite transition. Unchanged denotes cases with no change in correctness status. Net Gain is computed as Corrected minus Regressed.

Scale	**4B**				**9B**			
Benchmark	**Corrected**	**Regressed**	**Unchanged**	**Net Gain**	**Corrected**	**Regressed**	**Unchanged**	**Net Gain**
MedCaseReasoning	87	27	783	**+60**	133	55	709	**+78**
MedQA	76	17	1,180	**+59**	110	22	1,141	**+88**
MIMIC-CDM-FI	20	0	980	**+20**	42	6	952	**+36**
ER-Reason	23	0	977	**+23**	40	10	950	**+30**

As shown in Table 2, *MedGuideX* consistently produces positive net gains across all benchmarks and both model scales. At the 4B scale, *MedGuideX* corrects substantially more backbone errors than it introduces regressions, yielding net gains of 60 on MedCaseReasoning, 59 on MedQA, 20 on MIMIC-CDM-FI, and 23 on ER-Reason. The same trend holds at the 9B scale. These results indicate

Table 3: Clinical reasoning evaluation comparing *MedGuideX*-9B and GPT-5.0 (%). Reasoning recall is the fraction of clinician-authored reference reasoning steps recovered in the model rationale on MedCaseReasoning, scored by an LLM judge. Physician pairwise win rates report blinded non-tie preference rates of *MedGuideX*-9B over GPT-5.0 on 30 case annotations.

Metric	**Score**
Reasoning Recall	GPT-5.0: 50.78 *MedGuideX*-9B: **62.77**
Physician Pairwise Win Rate: MedGuideX-9B over GPT-5.0	Faithfulness: **85.00** Validity: **79.31** Completeness: **92.31** Clarity: **51.52** Overall: **76.86**

that guideline-derived post-training improves performance primarily by correcting backbone errors rather than by introducing unpredictable answer shifts.

4.4 Clinical Reasoning Process Evaluation

Final-answer accuracy does not directly measure whether a model reaches its answer through clinically sound reasoning. We therefore evaluate rationale quality using two complementary protocols: automatic reasoning recall against clinician-authored reference traces and a blinded physician preference study.

Reasoning recall. MedCaseReasoning includes clinician-authored reasoning traces for each case, which allows us to measure *reasoning recall*: the fraction of salient reference reasoning steps recovered in the model generated rationale, scored by an LLM judge following the paper's protocol (Appendix D). As reported in Table 3, *MedGuideX*-9B achieves a reasoning recall of 62.77%, compared with 50.78% for GPT-5.0. This suggests that *MedGuideX*-9B recovers more clinically relevant intermediate reasoning steps, indicating that guideline-derived post-training improves not only answer prediction but also the coverage of clinician-aligned reasoning.

Blinded physician evaluation. We conduct a blinded pairwise physician evaluation on 30 MedCaseReasoning cases where both *MedGuideX*-9B and GPT-5.0 produce the correct final diagnosis. Two practicing physicians each annotate 20 cases. For each case, the physician compares two anonymized responses shown in randomized order and judges four dimensions: evidence faithfulness, reasoning validity, reasoning completeness, and reasoning clarity. We treat *Same* as a tie and report non-tie win rates of *MedGuideX*-9B over GPT-5.0.

As shown in Table 3, *MedGuideX*-9B is preferred in faithfulness (85.00%), validity (79.31%), completeness (92.31%), and clarity (51.52%), with an overall non-tie win rate of 76.86%. The largest advantages appear in completeness and faithfulness, suggesting that *MedGuideX*-9B provides more complete reasoning and better grounding in case evidence. The clarity score is close to parity, indicating that *MedGuideX*-9B maintains comparable presentation quality while improving the clinically substantive dimensions of rationale quality. These results show that the proposed training improves reasoning quality in ways that are visible to practicing physicians and not captured by final-answer accuracy alone.

5 Conclusion

Inspired by clinical practice, we introduced *MedGuideX*, a medical LLM trained to internalize CPG decision logic for clinical reasoning. Instead of treating guidelines merely as unstructured textual knowledge, we transform high-quality guidelines into executable, structured training supervision. This supervision includes factual and counterfactual QA instances that expose models to evidence-based clinical decision rules. Through SFT and RL post-training, experiments on four benchmarks show that *MedGuideX* substantially improves over the base model and outperforms similarly sized medical LLMs. Further human evaluations demonstrate that *MedGuideX* improves diagnostic accuracy and generates more clinically grounded, coherent, and reliable rationales. These results highlight the value of moving beyond surface-level medical knowledge toward structured, executable clinical decision logic. They also suggest that CPGs can provide scalable supervision for reliable medical LLMs that reason more consistently with evidence-based clinical practice.

REFERENCES

Anthropic. Claude haiku 4.5 [large language model]. https://claude.ai/, 2025. Released October 15, 2025.

Judith L Bowen. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21):2217–2225, 2006.

Lang Cao. Graphreason: Enhancing reasoning capabilities of large language models through a graph-based verification approach. In *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ ACL 2024)*, pp. 1–12, 2024.

Lang Cao, Jingxian Xu, Hanbing Liu, Jinyu Wang, Mengyu Zhou, Haoyu Dong, Shi Han, and Dongmei Zhang. Formula-r1: Incentivizing llm reasoning over complex tables with numerical computation via formula-driven reinforcement learning. *arXiv preprint arXiv:2505.23667*, 2025.

Lang Cao, Qingyu Chen, and Yue Guo. Ehr-rag: Bridging long-horizon structured electronic health records and large language models via enhanced retrieval-augmented generation. *arXiv preprint arXiv:2601.21340*, 2026.

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.

Yuefei Chen, Vivek K Singh, Jing Ma, and Ruxiang Tang. Counterbench: A benchmark for counterfactuals reasoning in large language models. *arXiv preprint arXiv:2502.11008*, 2025.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, François Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. A modular approach for clinical slms driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19352–19374, 2025.

Ruiqi Deng, Geoffrey Martin, Tony Wang, Gongbo Zhang, Yi Liu, Chunhua Weng, Yanshan Wang, Justin F Rousseau, and Yifan Peng. Cpgprompt: Translating clinical guidelines into llm-executable decision support. *arXiv preprint arXiv:2601.03475*, 2026.

Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Marta Gonzalez-Mallo, et al. The aloe family recipe for open and specialized healthcare llms. *arXiv preprint arXiv:2505.04388*, 2025.

Boyang Gu, Hongjian Zhou, Bradley Max Segal, Jinge Wu, Zeyu Cao, Hantao Zhong, Lei Clifton, Fenglin Liu, and David A Clifton. Clinical-r1: Empowering large language models for faithful and comprehensive reasoning with clinical objective relative policy optimization. *arXiv preprint arXiv:2512.00601*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, and Jiawei Han. Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *arXiv preprint arXiv:2503.00223*, 2025a.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025b.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

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):6421, 2021.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, pp. 5848–5864, 2024.

Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. *arXiv preprint arXiv:2510.04284*, 2025.

Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*, 2023a.

Bingxuan Li, Simo Du, and Yue Guo. Joint optimization of reasoning and dual-memory for self-learning diagnostic agent. *arXiv preprint arXiv:2604.07269*, 2026.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023b.

Wenliang Li, Rui Yan, Xu Zhang, Li Chen, Hongji Zhu, Jing Zhao, Junjun Li, Mengru Li, Wei Cao, Zihang Jiang, et al. Macd: Multi-agent clinical diagnosis with self-learned knowledge for llm. *arXiv preprint arXiv:2509.20067*, 2025.

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 *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Nikita Mehandru, Niloufar Golchini, David Bamman, Travis Zack, Melanie F Molina, and Ahmed Alaa. Er-reason: A benchmark dataset for llm-based clinical reasoning in the emergency room. *arXiv preprint arXiv:2505.22919*, 2025.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Mathieu Nendaz and Arnaud Perrier. Diagnostic errors and flaws in clinical reasoning: mechanisms and prevention in practice. *Swiss medical weekly*, 142(4344):w13706–w13706, 2012.

David Oniani, Xizhi Wu, Shyam Visweswaran, Sumit Kapoor, Shravan Kooragayalu, Katelyn Polanska, and Yanshan Wang. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pp. 694–702. IEEE, 2024.

Marc Cicero Schubert, Stella Soyka, Wolfgang Wick, and Varun Venkataramani. Guideline-incorporated large language model-driven evaluation of medical records using medcheckllm. *JMIR Formative Research*, 9:e53335, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Harold C Sox, Michael C Higgins, Douglas K Owens, and Gillian Sanders Schmidler. *Medical decision making*. John Wiley & Sons, 2024.

Michael Staniek, Artem Sokolov, and Stefan Riezler. Training and evaluation of guideline-based medical reasoning in llms. *arXiv preprint arXiv:2512.03838*, 2025.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *Advances in Neural Information Processing Systems*, 37:7821–7846, 2024.

Anni Tziakouri and Filippo Menolascina. Reinforcement learning for clinical reasoning: Aligning llms with acr imaging appropriateness criteria. *arXiv preprint arXiv:2510.05194*, 2025.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Aniket Vashishtha, Qirun Dai, Hongyuan Mei, Amit Sharma, Chenhao Tan, and Hao Peng. Executable counterfactuals: Improving llms' causal reasoning through code. *arXiv preprint arXiv:2510.01539*, 2025.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025a.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J Tao, Min Woo Sun, Alejandro Lozano, and James Zou. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*, 2025b.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rui Yang, Han Zhong, Jiawei Xu, Amy Zhang, Chongjie Zhang, Lei Han, and Tong Zhang. Towards robust offline reinforcement learning under diverse data corruption. In *International Conference on Learning Representations*, volume 2024, pp. 15512–15543, 2024.

Gwangpyo Yoo and Honguk Woo. Model risk-sensitive offline reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhiwen You, Xi Chen, Aniket Vashishtha, Simo Du, Gabriel Erion-Barner, Hongyuan Mei, Hao Peng, and Yue Guo. Improving clinical diagnosis with counterfactual multi-agent reasoning. *arXiv preprint arXiv:2603.27820*, 2026.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

CONTENTS OF APPENDIX

This appendix provides additional details supporting the main paper.

- **A. Limitations** summarizes the current limitations of *MedGuideX* and its intended research-use scope.
- **B. Ethical Statement** discusses data privacy, research use, and clinical safety considerations.
- **C. Data Preparation Details** describes guideline curation, executable transformation, and QA synthesis statistics.
- **D. Benchmark Details and Examples** details the evaluation datasets, scoring protocols, licenses, and representative examples.
- **E. Training Details** reports SFT and RL settings for training *MedGuideX*.
- **F. Training Strategy Ablation** analyzes the effects of data composition and training phases.
- **G. Human Study Details** describes the physician evaluation design, annotation protocol, scoring, and questionnaire.
- **H. Details of Guideline-Based Baselines** explains the implementation of retrieval, in-context, fine-tuning, and RL guideline baselines.
- **I. Case Study** presents qualitative diagnostic case studies comparing *MedGuideX* with baseline models.
- **J. AI Use Statement** describes AI assistance in writing, literature search, and synthetic data generation.

A LIMITATIONS

While our work demonstrates promising results, it still has limitations. *MedGuideX* should be viewed as a research system for supporting studies of medical reasoning rather than as a substitute for clinical professionals. Further validation, safety evaluation, regulatory review, and integration with clinical workflows are needed before deployment in production environments or real-world clinical use.

B ETHICAL STATEMENT

This work aims to improve clinical reasoning in medical LLMs by transforming publicly available clinical practice guidelines into structured training supervision. All guideline sources used in this study are publicly accessible and do not contain private patient information. The factual and counterfactual QA instances are synthetically generated from executable guideline functions and are not derived from identifiable patient records.

We acknowledge that medical LLMs may produce incorrect, incomplete, or potentially harmful recommendations, even when trained on guideline-derived data. Therefore, *MedGuideX* is intended for research use only and should not be used as a substitute for professional medical judgment, clinical diagnosis, or treatment decisions. Any deployment in real-world healthcare settings would require rigorous clinical validation, human oversight, bias and safety evaluation, and compliance with applicable medical regulations.

C DATA PREPARATION DETAILS

C.1 GUIDELINE CURATION

Our initial guideline source is an open CPG collection containing 37,970 documents. After curation, we obtain a filtered subset of 841 CPGs, including 607 from CDC and 234 from PubMed.

C.2 EXECUTABLE TRANSFORMATION

Starting from 841 curated documents, we split the corpus into 3,127 recommendation-oriented chunks using a soft limit of 4,500 words per chunk, with at most 4 chunks per document and

Table 4: Data pipeline statistics on the guideline corpus. Each stage discards instances that fail automated checks.

Pipeline stage	Count
US-only guideline documents	841
CDC	607
PubMed	234
Guideline chunks	3,127
Recommendation candidates	4,750
Validated actionable recommendations	3,796
Validated decision trees	2,800
Executable guideline functions	2,793
Factual QA samples	4,963
Counterfactual QA candidates	5,028

2 recommendations per chunk. The LLM extractor produces 4,750 recommendation candidates. After validation, we retain 3,796 usable recommendations, discarding 827 candidates that lack a concrete condition-action structure and removing 127 near-duplicates with highly similar populations, conditions, and actions.

Each of the 3,796 retained recommendations is converted into a decision tree T_f specifying input variables, decision conditions, and final outputs. After LLM-based validation for completeness, condition clarity, branch coverage, and guideline support, 2,800 trees pass validation.

The 2,800 validated trees are compiled into executable Python functions that map input variables X to guideline outputs $f(X)$. We check each function for syntactic correctness, executability on sampled inputs, and consistency with the original decision tree. In total, 2,793 functions pass these checks and form our validated executable guideline knowledge base.

C.3 QA Synthesis

At the data level, executing f before and after intervention yields 7,993 interventional scenarios. Of these, 7,205 produce changed outcomes, while 788 unchanged-output cases are discarded. After applying the same balancing rule used for factual QA, we retain 5,028 counterfactual candidates. Each sample is paired with a GPT-5.4-mini generated reasoning trace that verbalizes the abduction, intervention, and prediction steps over the executable function f. Overall data pipeline statistics are shown in Table 4.

D Benchmark Details and Examples

We evaluate *MedGuideX* on four external medical reasoning benchmarks. All four are used only for evaluation and are disjoint from the executable guideline corpus used to train *MedGuideX*.

- **MedQA** (Jin et al., 2021): A multiple-choice medical exam benchmark based on United States Medical Licensing Examination (USMLE)-style questions. It evaluates whether executable guideline-grounded training transfers to exam-style medical knowledge. We report multiple-choice accuracy on 1,273 test questions.
- **MedCaseReasoning** (Wu et al., 2025b): A long-form open diagnostic reasoning benchmark based on open-access clinical case reports from the PubMed Central Open Access Subset. It evaluates case-based differential diagnosis. We report diagnostic accuracy on the 897 cases in the test split.
- **MIMIC-CDM-FI** (Hager et al., 2024): A full-information open clinical decision-making benchmark derived from MIMIC-IV, which is based on electronic health records from Beth Israel Deaconess Medical Center. It evaluates full-information clinical decision making. We report diagnostic accuracy on a randomly sampled subset of 1,000 cases.
- **ER-Reason** (Mehandru et al., 2025): An emergency-department open diagnosis prediction benchmark derived from patient records at a large academic medical center. It evaluates emergency-room diagnosis prediction. We report diagnostic accuracy on a randomly sampled subset of 1,000 cases.

Unless otherwise stated, we use greedy decoding with a 1,024-token generation budget, disable Qwen thinking mode, and use `gpt-5.4-mini` for all LLM judges. For each benchmark, we describe the task, split, and evaluation protocol, and provide one representative example.

MedQA is a USMLE-style multiple-choice medical QA benchmark. We evaluate on the US test split of 1,273 questions. Given a question stem and five options A through E, the model selects the single best answer after a chain-of-thought reasoning step. Answer extraction is fully deterministic: a regex-based parser strips reasoning blocks and extracts the final option letter or option text, with no LLM judge involved. We report accuracy as the fraction of correctly answered questions.

MedQA: USMLE-style multiple-choice QA

SYSTEM
You are a medical expert. Answer the multiple-choice question, thinking step by step before choosing one option from A to E.

USER
Context. A 65-year-old man with fever, productive cough, pleuritic chest pain, and right lower-lobe consolidation on chest x-ray. Gram stain shows lancet-shaped gram-positive diplococci.
Question. Which option is the single best answer?
Options. A: *Mycoplasma pneumoniae* B: *Streptococcus pneumoniae* C: *Pseudomonas aeruginosa* D: *Pneumocystis jirovecii* E: *Histoplasma capsulatum*

Reference diagnosis
B

Evaluation
The deterministic parser extracts option B from the model output and matches it against the gold answer B. The prediction is counted as correct.

MedCaseReasoning evaluates long-form diagnostic reasoning from open-access clinical case reports. Each example provides a clinical case prompt, a clinician-authored reasoning trace, and a final diagnosis. We evaluate on the official test split of 897 cases and use the official prompt template, which requests reasoning inside `<think>` tags and the diagnosis inside `<answer>` tags. We report 1-pass diagnostic accuracy under the official LLM-as-judge protocol, in which the judge compares the predicted and gold diagnoses and counts synonyms, abbreviations, and close medical paraphrases as correct. We additionally report reasoning recall, computed by a separate judge that checks, for each clinician-authored reasoning step, whether the model's reasoning explicitly or implicitly covers it.

MedCaseReasoning: long-form case diagnosis

USER
Please reason through the following clinical case and predict the final diagnosis, returning reasoning inside `<think>` tags and the diagnosis inside `<answer>` tags.
Clinical case. A middle-aged patient presents with severe episodic headaches, palpitations, sweating, and paroxysmal hypertension. Abdominal imaging reveals an adrenal mass. Plasma metanephrines are markedly elevated.

Reference diagnosis
Pheochromocytoma

Evaluation
The judge compares the predicted diagnosis against the gold diagnosis pheochromocytoma and finds them equivalent. Reasoning recall is computed over the clinician-authored reasoning steps; in this case all reference steps are covered, giving a recall of 1.0.

MIMIC-CDM-FI is a full-information clinical decision-making benchmark derived from the MIMIC-IV-Ext CDM dataset. We evaluate on a fixed 1,000-case test split balanced across four acute abdominal pathologies, with 250 cases each of appendicitis, cholecystitis, pancreatitis, and diverticulitis. The model receives all relevant patient information upfront, including history, physical examination, selected laboratory results, and abdominal imaging reports, and is asked to output a single final diagnosis of the most severe pathology with no further explanation. Predictions are scored by an LLM-as-judge that identifies the primary acute abdominal diagnosis. Synonyms, spelling

variants, abbreviations, and clinically equivalent paraphrases count as correct; comorbidities or full billing-code matches are not required.

MIMIC-CDM-FI: full-information clinical decision making

SYSTEM
You are a medical AI assistant. Based on the provided information, give a single final diagnosis of the most severe pathology, with no further information.

USER
Provide the most likely final diagnosis of the following patient.
History. Acute right lower quadrant abdominal pain, nausea, anorexia, low-grade fever.
Examination. Tenderness at McBurney's point with guarding.
Laboratory. Elevated white blood cell count and C-reactive protein.
Imaging. CT abdomen shows a dilated appendix with wall thickening and periappendiceal fat stranding.

Reference diagnosis
Acute appendicitis

Evaluation
The judge identifies the model's response as a match for the gold diagnosis acute appendicitis. The prediction is counted as correct.

ER-Reason evaluates clinical reasoning in emergency-department settings; we use Task 4, final ED diagnosis prediction. We evaluate on a deterministic $1{,}000$-record split. Each record provides age, sex, chief complaint, the current ED presentation, and a set of clinical notes spanning discharge summary, progress notes, history and physical, imaging, and consult. The model receives all notes as context and outputs a single ED diagnosis in free text. Our primary metric is LLM-judge accuracy, with the judge counting synonyms, abbreviations, wording differences, and clinically equivalent specificity as correct. The official Task 4 evaluator also provides a CMS ICD-10-CM crosswalk that yields exact match, normalized match, token F1, clinical-cluster, body-system, ICD, and HCC accuracy. We do not adopt the crosswalk-based HCC accuracy as our primary metric because the crosswalk maps only about 30% of the gold and predicted diagnoses to an HCC category. The resulting score is computed over a small and non-representative subset of cases, whereas LLM-judge accuracy scores every case and better reflects clinically correct predictions.

ER-Reason (Task 4): emergency-department diagnosis

SYSTEM
You are an experienced ED physician. Predict the most likely diagnosis for the current ED visit, using all available clinical notes as full information.

USER
ER-Reason Task 4: Final Diagnosis. Age 72, Female. Chief complaint: shortness of breath.
Clinical notes. History of heart failure; worsening dyspnea, orthopnea, bilateral leg edema, elevated BNP, pulmonary vascular congestion on chest x-ray.
Current ED presentation. Tachypneic with hypoxia and bibasilar crackles; no fever or focal infiltrate.
Output a single CMS-standardized diagnosis as text, with no explanation.

Reference diagnosis
Acute decompensated heart failure

Evaluation
The judge compares the model's prediction acute decompensated heart failure against the gold acute congestive heart failure and counts them as clinically equivalent. The prediction is counted as correct.

Licenses and access. We confirm that all four evaluation benchmarks are used in accordance with their original licenses, and only for non-commercial academic research. MedQA is released under the MIT license via the official GitHub repository.[3] MedCaseReasoning's code is released under the MIT license and its dataset is released under CC-BY 4.0, derived from the PubMed Central Open Access Subset.[4,5] MIMIC-CDM-FI is distributed under the PhysioNet Credentialed Health Data License 1.5.0 with an accompanying Data Use Agreement, and access requires completion of the CITI Data

[3] https://github.com/jind11/MedQA
[4] https://github.com/kevinwu23/Stanford-MedCaseReasoning
[5] https://huggingface.co/datasets/zou-lab/MedCaseReasoning

Table 5: Summary of the four evaluation benchmarks. # Eval is the number of evaluated examples. MedQA uses a fully deterministic answer parser with no LLM judge; the other three use an LLM-as-judge (`gpt-5.4-mini`, temperature 0) for semantic diagnosis matching. ER-Reason also has an official CMS ICD-10-CM crosswalk evaluator available, but it is not used as the primary metric (see Appendix D).

Benchmark	Split	# Eval	Scoring
MedQA	US test	1,273	parser
MedCaseReasoning	test	897	LLM judge
MIMIC-CDM-FI	test_1000	1,000	LLM judge
ER-Reason	test_1000	1,000	LLM judge

or Specimens Only Research training; we accessed it as credentialed PhysioNet users and use it only for evaluation in this study.[6] ER-Reason is distributed under the more restrictive PhysioNet Contributor Review Health Data License 1.5.0, which additionally requires per-study review by the dataset contributors; we accessed it under this credentialed and reviewed access process and use it only for evaluation.[7] We do not redistribute any patient-level records from MIMIC-CDM-FI or ER-Reason, and all evaluation outputs reported in this paper are aggregate metrics rather than raw clinical content.

Summary. Table 5 summarizes the four benchmarks. MedQA isolates broad exam-style medical knowledge with deterministic scoring; MedCaseReasoning targets long-form differential diagnosis with both accuracy and reasoning-recall judging; MIMIC-CDM-FI measures full-information acute abdominal decision making on a class-balanced split; and ER-Reason measures ED diagnosis prediction with both semantic and code-crosswalk evaluation.

E Training Details

Training setup. We train *MedGuideX*-4B and *MedGuideX*-9B from Qwen3.5-4B and Qwen3.5-9B, respectively. Both models use the same post-training recipe: supervised fine-tuning on the guideline-derived factual and counterfactual QA mixture, followed by GRPO on factual guideline QA prompts. Unless otherwise stated, the details below describe the 9B run; the 4B run follows the same data construction and optimization setup with the corresponding Qwen3.5-4B backbone.

Supervised fine-tuning. The SFT stage is designed to teach the model how to express the executable guideline logic in natural-language reasoning. We train on the balanced mixture of factual QA and counterfactual QA examples described in Section 3.2. For factual examples, the target response contains the guideline-prescribed output and a rationale verbalizing the executed path $\pi_f(X)$. For counterfactual examples, the target response follows the abduction–intervention–prediction structure, requiring the model to infer hidden variables, apply the intervention, and predict the new guideline output. This stage therefore serves as a cold start for both the answer format and the reasoning pattern used in later RL.

All SFT runs use LoRA adaptation on all linear layers, with rank 16 and $\alpha = 32$, bfloat16 precision, and gradient checkpointing. We train with verl (Sheng et al., 2025) for 5 epochs, using a learning rate of 1×10^{-5}, cosine warmup ratio 0.03, maximum sequence length 2048, and global batch size 64. For the 9B model, the SFT stage runs for approximately 16 hours on 8× RTX 5090 GPUs, and the resulting checkpoint is used to initialize the RL stage.

Reinforcement learning. After SFT, we apply GRPO (Shao et al., 2024) using the same verl training stack. The final *MedGuideX* configuration uses factual RL, selected by the ablation study in Section F. The RL prompts reuse the 4,963 factual prompts from the guideline-derived SFT corpus. Since all four evaluation benchmarks are external to the executable guideline corpus, this reuse does not introduce benchmark contamination.

[6] https://physionet.org/content/mimic-iv-ext-cdm/
[7] https://physionet.org/content/er-reason/1.0.0/

During RL, each prompt is sampled with multiple rollouts and scored by the reward function described in Section 3.3. The reward first enforces the required reasoning-before-answer format, then checks whether the parsed final guideline output matches the oracle output produced by the executable function. For counterfactual RL ablations, the same reward implementation additionally verifies hidden-variable recovery and executable counterfactual consistency, but this is not the final configuration used for *MedGuideX* in Table 1.

The RL stage is initialized from the SFT checkpoint and trained for one epoch with learning rate 5×10^{-6}. We use a maximum prompt length of 1536, maximum response length of 1024, 24 rollouts per prompt, sampling temperature 1.0, top-$p = 1.0$, one PPO update epoch, entropy coefficient 0.0, and KL coefficient 0.005 with the low-variance KL estimator. We evaluate and save checkpoints every 10 training steps, selecting the final checkpoint according to performance on the held-out validation splits. For the 9B model, the RL stage runs for approximately 33 hours on $8\times$ RTX 5090 GPUs, and the resulting checkpoint is final MedGuideX-9B checkpoint.

F Training Strategy Ablation

We ablate two components of the 9B training pipeline: the *data type* used in each phase, including factual QA, counterfactual QA, or a balanced mixture of both, and the *training phase*, including SFT, RL, or SFT followed by RL. Figure 2 reports accuracy changes relative to the Qwen3.5-9B backbone on MedCaseReasoning and MedQA. We also include a *continued pretraining on raw guidelines* baseline that uses the same raw guideline text without generated QA supervision. The goal of this ablation is to identify the most effective way to combine data composition and training phases for internalizing decision logic from executable clinical guidelines into LLMs for clinical reasoning.

The results show that exposure to guideline text alone is insufficient: continued pretraining on raw CPG text substantially decreases performance, despite being a common strategy in prior medical LLM training (Chen et al., 2023). In contrast, most configurations using generated QA supervision improve over the backbone, suggesting that structured factual and counterfactual QA examples are important for transferring guideline knowledge into reasoning behavior.

Among SFT-only settings, the balanced mixture performs best on both benchmarks, indicating that factual and counterfactual examples provide complementary supervision during supervised training. For RL-only settings, factual QA gives the strongest gains, suggesting that factual examples provide a more stable optimization signal for RL. The best overall configuration applies mixed-data SFT followed by factual RL. This configuration corresponds to *MedGuideX*-9B in Table 1 and supports our final training design. We hypothesize that factual reasoning is more closely aligned with downstream clinical reasoning benchmarks, making it a more effective final-stage RL objective before inference.

G Human Study Details

Study design. We conduct a blinded pairwise comparison study to assess the clinical reasoning quality of *MedGuideX*-9B against GPT-5.0. The study covers 30 clinical cases sampled from the subset of the MedCaseReasoning test split on which both models produce a correct final diagnosis, so that the comparison isolates rationale quality from answer correctness. For each case, an annotator is shown the case prompt, the ground-truth diagnosis, and two anonymized responses, and judges which response is better along four reasoning dimensions. Model identities are hidden, and the left/right presentation order of the two responses is independently randomized per case so that the annotator cannot infer which system produced which response.

Annotators and case allocation. Annotations are provided by two practicing physicians. Each physician annotates 20 cases, with 10 cases shared between them. Concretely, physician 1 annotates cases 1–20 and physician 2 annotates cases 11–30, so that cases 11–20 are double annotated. This design covers all 30 cases with 40 total annotations while keeping per-annotator load manageable, and the shared 10 cases provide a within-study measure of inter-annotator agreement. Annotators are not told the identity, size, or training procedure of either model.

Evaluation dimensions. The four dimensions and the instructions shown to each annotator are as follows.

Case 010 — Clinical question

A 29-year-old Indigenous Australian man presented with severe (8/10) epigastric pain radiating to his back. History: asthma, type 2 diabetes, obesity, hypertriglyceridemia, prior hypertriglyceridemia-induced pancreatitis three years earlier (triglycerides 39 mmol/L). Current smoker, denies alcohol. On arrival: T 36.5°C, HR 64, BP 138/83, RR 11, SpO_2 99%, clinically dehydrated. Post-resuscitation labs: triglycerides 24.1 mmol/L (<1.70), cholesterol 8.4 mmol/L, lipase 119 U/L (8–78), WBC 12.3×10^9/L, corrected calcium 2.29 mmol/L, creatinine 68 µmol/L, glucose 21.3 mmol/L, no ketonuria. Contrast-enhanced abdominal CT at 24 h: fatty liver, no features of pancreatitis. *[. . . full prompt shown to annotator. . .]*

Ground-truth final diagnosis: hyperlipemic abdominal crisis

Response A

Reasoning: The patient has classic features of pancreatitis (severe epigastric pain radiating to the back, hypertriglyceridemia, prior episode), but the lipase is only mildly elevated and the CT at 24 h shows no pancreatic changes. Acute pancreatitis would typically show lipase 3–10× upper limit and imaging findings by this time. However, lipemia can interfere with lipase assays and the classic presentation plus severe hypertriglyceridemia still favor an underlying pancreatitis process. *[. . .]*

Final diagnosis: Acute pancreatitis secondary to hypertriglyceridemia

Response B

Reasoning: Multiple pancreatitis risk factors are present, but lipase is only mildly elevated and CT at 24 h shows no features of pancreatitis. In a patient with severe hypertriglyceridemia, severe abdominal pain without imaging or biochemical evidence of pancreatitis is most consistent with a hyperlipemic abdominal crisis, a recognized entity that mimics pancreatitis clinically. *[. . .]*

Final diagnosis: Hyperlipemic abdominal crisis

Case 010 — Pairwise judgments *Select Response A, Response B, or Same on each dimension.*

Dimension	Judgment			Comment
Evidence Faithfulness	☐ Response A	☐ Response B	☐ Same	*Optional elaboration*
Reasoning Validity	☐ Response A	☐ Response B	☐ Same	*Optional elaboration*
Reasoning Completeness	☐ Response A	☐ Response B	☐ Same	*Optional elaboration*
Reasoning Clarity	☐ Response A	☐ Response B	☐ Same	*Optional elaboration*

Figure 3: Layout of one case page in the blinded physician questionnaire. Each page shows the clinical question and the ground-truth final diagnosis, two blinded responses (one from *MedGuideX*-9B and one from GPT-5.0, presentation order independently randomized per case), and four mandatory single-choice questions, one per evaluation dimension. Annotators may select Response A, Response B, or Same, and can optionally provide free-text elaboration for each judgment. The ground-truth diagnosis is shown to the annotator so that judgments compare reasoning quality against a known correct answer; sample selection is further restricted to cases on which both models produce a correct final diagnosis. Case and response text are abbreviated here for space.

Table 6: Raw judgment counts from the blinded pairwise physician evaluation. Each cell is the number of annotations in which a given outcome was chosen. The total per dimension is 40, corresponding to 30 cases with 10 double-annotated cases. The non-tie win rates in Table 3 are computed by excluding the Same column from the denominator.

Dimension	*MedGuideX*-9B	GPT-5.0	Same
Evidence Faithfulness	17	3	20
Reasoning Validity	23	6	11
Reasoning Completeness	36	3	1
Reasoning Clarity	17	16	7
Overall (pooled)	93	28	39

- **Evidence Faithfulness.** Which response is more faithful to the evidence provided in the case? Prefer the response that grounds its reasoning in the stated history, symptoms, labs, imaging, and pathology, and penalize unsupported claims, invented findings, or conclusions that go beyond the provided evidence.
- **Reasoning Validity.** Which response has more valid clinical reasoning? Prefer the response whose diagnostic logic is medically sound, internally consistent, and appropriate for the task, and penalize incorrect causal links, medically implausible interpretations, or contradictions.
- **Reasoning Completeness.** Which response provides a more complete reasoning process? Prefer the response that covers the key positive and negative evidence, necessary intermediate steps, and important differential considerations, and penalize responses that skip critical evidence or jump to a diagnosis without adequate justification.
- **Reasoning Clarity.** Which response is more concise and focused while still being clinically useful? Prefer responses that are clear, organized, and free of unnecessary repetition or irrelevant detail, without rewarding brevity that omits important reasoning.

Annotation protocol. The study is administered as a structured questionnaire. Each case occupies a separate page containing, in order, the clinical *Question*, *Response A*, *Response B*, and four mandatory single-choice questions, one per evaluation dimension. For each dimension the annotator selects Response A, Response B, or Same. Optional free-text elaboration is allowed but not required. Responses cannot be edited after submission, and no annotator email or identifying information is collected.

Scoring. After collection, the blinded A/B labels are decoded back to model identities using a held-out mapping that is never exposed during annotation. For each dimension we report the *non-tie win rate* of *MedGuideX*-9B over GPT-5.0, defined as the number of annotations in which *MedGuideX*-9B is preferred divided by the total number of non-tie annotations,

$$\text{Win Rate} = \frac{n_{\textit{MedGuideX}\text{-9B}}}{n_{\textit{MedGuideX}\text{-9B}} + n_{\text{GPT-5.0}}},$$

where $n_{\textit{MedGuideX}\text{-9B}}$ and $n_{\text{GPT-5.0}}$ are the counts of *MedGuideX*-9B-preferred and GPT-5.0-preferred judgments, respectively. Same judgments are reported but excluded from the denominator, which is standard for pairwise model comparisons where ties carry no signal about relative quality. The aggregate *Overall* win rate in Table 3 is computed by pooling all non-tie judgments across the four dimensions and all 40 annotations.

Per-dimension judgment counts. Table 6 reports the raw counts underlying the win rates in Table 3. *MedGuideX*-9B is preferred on every dimension, with the strongest margins on reasoning completeness (36 vs 3) and evidence faithfulness (17 vs 3). Reasoning clarity is the most contested dimension, with 17 preferences for *MedGuideX*-9B, 16 for GPT-5.0, and 7 ties.

Questionnaire example. Figure 3 shows the layout of a representative case page as presented to the annotator. The case prompt and the two candidate responses are reproduced in abbreviated form; in the actual questionnaire the full untruncated text is shown.

H Details of Guideline-Based Baselines

Retrieval-Augmented Guideline Prompting. We evaluate an inference-time retrieval baseline that uses the same raw US-only guideline chunks as external knowledge. We build a TF-IDF retriever over the 3,127 raw CPG chunks, using each chunk's topic, section path, and guideline text as the retrieval corpus. For each benchmark instance, we form a query from the test input only: the question stem and answer choices for MedQA, the case prompt for MedCaseReasoning, the patient information and clinical fields for MIMIC-CDM-FI, and the emergency-department presentation and available notes for ER-Reason. Gold labels are never used in retrieval. We retrieve the top-3 chunks, truncate each retrieved snippet to 700 characters, and prepend them to the original benchmark prompt with an instruction that the snippets may or may not be relevant. No model parameters are updated. This baseline obtains 31.88%, 74.71%, 82.50%, and 27.40% on the four benchmarks, with an average accuracy of 54.12%.

In-Context Guideline Demonstrations We also evaluate an in-context learning baseline that provides guideline-derived examples at inference time without parameter updates. We select three factual QA examples from the guideline-derived factual QA set. Each demonstration contains a patient/guideline query, a short reasoning trace, and the corresponding guideline output. These three demonstrations are fixed across all test instances and are prepended before the original benchmark prompt, with an instruction that they are demonstrations only and are not facts about the target case. Unlike RAG, this baseline does not retrieve case-specific guideline passages; it tests whether a small number of guideline-style demonstrations can induce the desired reasoning behavior in the base model. This baseline obtains 33.11%, 74.94%, 80.10%, and 28.20% on the four benchmarks, with an average accuracy of 54.08%.

CPGPrompt. We adapt CPGPrompt (Deng et al., 2026) as a train-free decision-tree prompting baseline. For each benchmark instance, we first retrieve the top-3 guideline recommendations from the validated recommendation corpus using TF-IDF retrieval. We then prompt Qwen3.5-9B to construct a compact CPGPrompt-style decision tree as a Python dictionary literal, including the selected recommendation ID, yes/no decision nodes, a final rationale, and a final answer. The generated tree is parsed, normalized, and executed locally by a deterministic Python traverser; the reached terminal action or top-level final answer is then scored under the same benchmark-specific evaluation protocol as the main experiments. This baseline does not use our precompiled executable functions and does not update model parameters. It obtains 14.16%, 59.15%, 67.60%, and 19.90% on the four benchmarks, with an average accuracy of 40.20%.

Fine-tuning with CPG. We include a text-only fine-tuning baseline that trains Qwen3.5-9B directly on raw CPG text. The training data are constructed from guideline chunks, where the model is asked to continue or reproduce the raw guideline content under a medical guideline system prompt. This baseline uses only unstructured guideline text and does not use extracted recommendations, decision trees, executable Python functions, generated factual/counterfactual QA pairs, or verifier-based rewards. Its goal is to test whether exposure to guideline text alone is sufficient to improve clinical reasoning. We evaluate this baseline on MedCaseReasoning and MedQA, where it obtains 28.87% and 65.04%, respectively. The results indicate that raw guideline text alone can lead to negative transfer compared with the Qwen3.5-9B backbone.

RL with CPG-Derived Process Rewards. Finally, we evaluate an RL baseline that uses CPG-derived process rewards instead of our executable-guideline supervision. This baseline starts from the same mixed factual/counterfactual SFT checkpoint and applies GRPO on factual guideline QA prompts. The reward function reuses the factual CoT format reward and final-answer reward, then adds an LLM-as-a-judge process reward: for each sampled response, we retrieve the most relevant validated guideline recommendation, using an exact CPG identifier when available and otherwise lexical overlap retrieval, and ask the judge whether the model's reasoning process is consistent with the retrieved guideline. The final reward is the sum of the format reward, answer reward, and process-consistency reward. Unlike *MedGuideX*, this baseline does not execute guideline functions to verify the model's reasoning path or counterfactual behavior. It obtains 37.68%, 75.73%, 83.40%, and 27.80% on the four benchmarks, with an average accuracy of 56.15%.

I CASE STUDY

We provide a controlled qualitative comparison on MedCaseReasoning, a long-form diagnostic reasoning benchmark in which each model receives the same clinical case and must produce a final diagnosis. To make the comparison directly auditable, we select three cases where *MedGuideX*, as shown in this section, is judged correct while both GPT-5.0 and MedReason-8B are judged incorrect. For each case, we preserve the complete clinical case, the complete raw model response, and the verbatim judge rationale for each prediction.

Green boxes indicate correct *MedGuideX* outputs, and red boxes indicate incorrect baseline outputs. These examples are not intended to replace aggregate evaluation; rather, they illustrate diagnostic distinctions captured by guideline-derived post-training but missed by the comparison models.

I.1 CASE 1: CENTRAL SEROUS CHORIORETINOPATHY

Task Prompt

Please reason through the following clinical case and predict the final diagnosis. Do not use tables or bullet points. Return your answer in the following format: ¡think¿ [Reasoning here] ¡/think¿ ¡answer¿ [Final diagnosis here] ¡/answer¿.

Full Clinical Case

A 42-year-old previously healthy woman presented with a 4-day history of blurring of vision and visual loss in her left eye, associated with mild ocular pain. She denied systemic symptoms, drug use, and had no significant family, medical, or ocular history, except for a type-A behavioral pattern. She had been diagnosed elsewhere with presumed unilateral Vogt–Koyanagi–Harada disease and referred for further evaluation prior to treatment.

On examination, best-corrected visual acuity was 20/20 in the right eye and 20/40 in the left. Cycloplegic refraction was +0.50 diopters in the right eye and -1.00 diopters in the left. Intraocular pressures were normal bilaterally. Slit-lamp biomicroscopy showed a clinically shallow anterior chamber in the left eye and +1 cells in the anterior vitreous; there was no anterior chamber inflammation. Funduscopic examination of the left eye revealed large areas of neurosensory retinal elevation consistent with subretinal fluid involving the peripapillary region and extending beyond the posterior pole; the right eye fundus was unremarkable.

Spectral-domain optical coherence tomography of the left eye confirmed serous detachment of the neurosensory retina in the macular, peripapillary, and juxtapapillary regions. Optical biometry demonstrated that the anterior chamber depth was 0.62 mm shallower in the left eye than in the right, and axial length was 0.16 mm shorter in the left eye.

Model Outputs.

MedGuideX (correct)

Prediction: Idiopathic central serous chorioretinopathy
Judge reason: Idiopathic central serous chorioretinopathy is the same diagnosis as central serous chorioretinopathy; the added idiopathic qualifier does not change the condition.

Raw response:
The patient is a 42-year-old woman with acute unilateral visual loss and blurring in the left eye, mild ocular pain, and a shallow anterior chamber. She was previously diagnosed with Vogt-Koyanagi-Harada disease, but the current presentation is atypical for VKH, which typically presents with bilateral uveitis, meningismus, and systemic symptoms such as hearing loss or skin changes. Here, the anterior chamber inflammation is minimal (only +1 cells), and there is no significant vitritis or posterior uveitis on exam. Instead, the key findings are a shallow anterior chamber, reduced axial length, and extensive serous retinal detachment involving the macula, peripapillary region, and beyond the posterior pole. The shallow anterior chamber and shorter axial length in the affected eye suggest a structural abnormality rather than an inflammatory process. These features are characteristic of a primary vitreoretinal disorder, specifically idiopathic central serous chorioretinopathy (CSC), which can present with serous retinal detachment and sometimes a shallow anterior chamber due to ciliary body effusion or altered aqueous dynamics. However, the combination of shallow anterior chamber, reduced axial length, and extensive serous detachment is more consistent with a diagnosis of idiopathic central serous chorioretinopathy with associated anterior chamber shallowing, possibly due to ciliary body effusion or a variant of CSC. Given the absence of significant uveitis and the presence of serous retinal detachment with shallow anterior chamber, the most likely diagnosis is idiopathic central serous chorioretinopathy.

¡answer¿
Idiopathic central serous chorioretinopathy
¡/answer¿

GPT-5.0 (wrong)

Prediction: Posterior scleritis
Judge reason: Posterior scleritis is a different inflammatory ocular condition and is not a synonym or close paraphrase of central serous chorioretinopathy.

Raw response:
Reasoning:
Unilateral serous retinal detachment with mild ocular pain, shallow anterior chamber, axial shortening, and a paradoxical myopic shift indicates anterior displacement from ciliochoroidal effusion, favoring posterior scleritis. Mild vitreous cells and peripapillary/macular

detachment support inflammation rather than CSC. Vogt–Koyanagi–Harada is unlikely given the unilateral involvement and lack of significant anterior chamber inflammation or systemic features.

¡answer¿
Posterior scleritis
¡/answer¿

MedReason-8B (wrong)

Prediction: Persistent fetal vasculature (PFV)
Judge reason: Persistent fetal vasculature (PFV) is a different retinal developmental anomaly and is not a synonym or close paraphrase of central serous chorioretinopathy.

Raw response:
Based on the provided clinical details, the patient presents with unilateral visual impairment and blurring of vision in the left eye, along with mild ocular pain. The key findings include:

1. **Shallow anterior chamber** in the left eye compared to the right.
2. **Subretinal fluid** in the left eye, particularly around the optic disc and extending into the macula.
3. Normal intraocular pressure bilaterally.
4. No anterior chamber inflammation observed during slit-lamp examination.
5. A history of being diagnosed with Vogt-Koyanagi-Harada (VKH) disease, which typically involves bilateral involvement and uveitis.

The presence of a shallow anterior chamber suggests a structural abnormality, possibly related to the anatomy of the eye. However, VKH disease is usually characterized by bilateral symptoms and inflammation, which are not evident here. The absence of anterior chamber inflammation and the unilateral nature of the presentation make VKH less likely.

Given these observations, the most plausible explanation for the unilateral subretinal fluid and shallow anterior chamber without inflammation could be a congenital anomaly such as persistent fetal vasculature (PFV), also known as persistant hyperplastic primary vitreous (PHPV). This condition can lead to structural abnormalities like a shallow anterior chamber and subretinal fluid due to improper development of the hyaloid artery and lens-related anomalies.

Therefore, the final diagnosis aligning with the clinical findings would be **persistent fetal vasculature (PFV)**.

I.2 Case 2: Behcet's Disease

Task Prompt

Please reason through the following clinical case and predict the final diagnosis. Do not use tables or bullet points. Return your answer in the following format: ¡think¿ [Reasoning here] ¡/think¿ ¡answer¿ [Final diagnosis here] ¡/answer¿.

Full Clinical Case

An 8-year-old boy of Brazilian descent with autism presented with an 8-month history of painful oral ulcers. He had a one-time perianal ulcer, folliculitis on his scalp and buttocks, intermittent fever, and bloody stools. He had no genital ulcers. He was initially treated as PFAPA with steroids, but symptoms recurred when steroids were tapered. Family history was noncontributory.

On examination, he had multiple large oral ulcers on the tongue, palate, and lips. Abdominal examination was normal, with no tenderness or masses; there were no perianal ulcers, arthritis, or other skin findings. An extensive workup for systemic lupus erythematosus, herpes simplex virus, Epstein–Barr virus, tuberculosis, cyclic neutropenia, primary immunodeficiencies, and periodic fever syndromes was unremarkable. Ophthalmological examination was negative, with no pathergy. Buccal biopsies showed chronic inflammation with granulation tissue, no granulomas, and grew Fusobacterium periodontium.

Prometheus IBD sgi testing demonstrated a pattern consistent with Crohn's disease (positive anti-A4-Fa2 IgG, anti-Fax IgG, ATG16L1 heterozygous, NKX2-3 heterozygous, ICAM-1, SAA 179, elevated CRP; ASCA and ANCA negative). Multiple endoscopies and colonoscopies with ileal intubation were negative. Capsule endoscopy, CT of the abdomen, and MRI enterography showed no abnormalities.

Despite treatment with steroids, colchicine, infliximab, and methotrexate–later escalated to high-dose infliximab–the patient's bloody stools worsened and he developed severe malnutrition (BMI z-score -2.38). At month 18, a fourth colonoscopy revealed large, round ulcers in the terminal ileum with granulation tissue and chronic inflammatory infiltrates, but no granulomas or cobblestoning.

Treatment was switched to ustekinumab, after which gastrointestinal symptoms improved. At month 19, he developed bilateral hip effusions and later presented with headaches, emesis, and fevers. A lumbar puncture showed an opening pressure of 39 cm H2O; brain MRI/MRA/MRV revealed extensive thrombosis of the superior sagittal sinus, straight sinus, and proximal right transverse sinus. A hypercoagulability panel was negative.

Model Outputs.

MedGuideX (correct)

Prediction: Behçet's disease
Judge reason: Exact match with the gold diagnosis.

Raw response:
The patient is an 8-year-old boy with a prolonged history of painful oral ulcers, bloody stools, and intermittent fever, initially misdiagnosed as PFAPA. Despite negative workup for common infectious and autoimmune causes, and negative endoscopic and imaging studies for Crohn's disease, he had a positive Prometheus IBD sgi test pattern consistent with Crohn's disease. However, the endoscopic findings were atypical: large round ulcers with granulation tissue and chronic inflammation but no granulomas or cobblestoning, and no ileal involvement on capsule endoscopy, CT, or MRI enterography. His symptoms worsened with multiple Crohn's-directed therapies, including infliximab,

and he developed severe malnutrition. The key turning point was the development of bilateral hip effusions and then cerebral venous sinus thrombosis (CVST) with extensive thrombosis of the superior sagittal, straight, and right transverse sinuses, despite a negative hypercoagulability panel. This presentation is highly suggestive of Behçet's disease, which can present with oral ulcers, gastrointestinal involvement (often ileal or colonic ulcers), and thrombotic complications such as CVST. The absence of genital ulcers does not exclude Behçet's disease, as it can be atypical in children. The negative ASCA and ANCA, along with the lack of granulomas, further supports a diagnosis of Behçet's disease rather than Crohn's disease. The response to ustekinumab, which is effective in Behçet's disease, also supports this diagnosis.

¡answer¿
Behçet's disease
¡/answer¿

GPT-5.0 (wrong)

Prediction: Crohn's disease
Judge reason: Crohn's disease is a different inflammatory bowel disease and not a synonym or direct equivalent of Behçet's disease.

Raw response:
Reasoning:
Severe, recurrent oral ulcers with bloody stools and later terminal ileal ulcers showing chronic inflammation, plus an IBD serologic pattern, point to Crohn's disease; symptoms improved with ustekinumab. Behçet disease is less likely given absent genital ulcers, negative pathergy and ophthalmologic exams, and GI findings ultimately localized to terminal ileum consistent with Crohn's. Cerebral venous sinus thrombosis is a known IBD complication.

¡answer¿
Crohn's disease
¡/answer¿

MedReason-8B (wrong)

Prediction: Crohn's disease
Judge reason: Crohn's disease is not the same diagnosis as Behçet's disease; they can overlap clinically but are distinct conditions.

Raw response:
Crohn's disease

I.3 Case 3: Infective Endocarditis with Renal Embolization

Task Prompt

Please reason through the following clinical case and predict the final diagnosis. Do not use tables or bullet points. Return your answer in the following format: ¡think¿ [Reasoning here] ¡/think¿ ¡answer¿ [Final diagnosis here] ¡/answer¿.

Full Clinical Case

A 52-year-old woman with known unrepaired Type II tricuspid atresia with d-transposition of the great arteries, hypoplastic right ventricle, unrestricted atrial and ventricular septal defects, and multi-level pulmonic stenosis presented with sudden-onset severe right upper quadrant abdominal pain without nausea, vomiting, diarrhea, or dysuria. Two months earlier, she had flu-like symptoms treated with doxycycline, and she reported poor dentition with no dental care for two years. On examination, her blood pressure was 90/54 mmHg, she was afebrile, heart rate was 80 bpm, and oxygen saturation was 79% on room air. A grade-2 systolic ejection murmur was heard at the left upper sternal border. She had cyanotic fingers bilaterally but no Osler nodes, Janeway lesions, or splinter hemorrhages. Laboratory studies showed hemoglobin 17.9 g/dL and C-reactive protein 2.2 mg/dL; lactate, white blood cell count, and creatinine were within reference ranges. Computed tomography of the abdomen demonstrated an acute right renal infarct; cerebral computed tomography was negative. She was anticoagulated with heparin.

Model Outputs.

MedGuideX (correct)

Prediction: Infective endocarditis with septic embolization to the right kidney
Judge reason: The prediction includes the gold diagnosis, infective endocarditis; the added septic embolization to the right kidney is a complication and does not change the core diagnosis.

Raw response:
The patient has a complex cyanotic congenital heart disease with unrepaired tricuspid atresia, d-transposition of the great arteries, hypoplastic right ventricle, unrestricted atrial and ventricular septal defects, and multi-level pulmonic stenosis. This anatomy results in right-to-left shunting and systemic cyanosis, as evidenced by her low oxygen saturation and cyanotic fingers. She presents with sudden severe right upper quadrant abdominal pain, and imaging reveals an acute right renal infarct. Her history includes recent doxycycline use for flu-like symptoms and poor dentition, which raises concern for a source of bacteremia. Although she lacks classic peripheral stigmata of infective endocarditis such as Osler nodes or Janeway lesions, the presence of a new renal infarct in the setting of cyanotic heart disease is highly suggestive of septic emboli from infective endocarditis. The murmur at the left upper sternal border may represent her baseline murmur from pulmonic stenosis, but the acute presentation with organ infarction points to embolic complications. Given her cyanotic heart

disease, she is at high risk for infective endocarditis due to turbulent flow and endothelial damage, and the renal infarct is a classic manifestation of septic emboli. Therefore, the most likely diagnosis is infective endocarditis with septic embolization to the kidney.

¡answer¿
Infective endocarditis with septic embolization to the right kidney

¡/answer¿

GPT-5.0 (wrong)

Prediction: Paradoxical systemic arterial embolism causing acute right renal infarction in cyanotic congenital heart disease (unrepaired tricuspid atresia with d-TGA
Judge reason: Predicted diagnosis is paradoxical systemic arterial embolism/renal infarction due to cyanotic congenital heart disease, which is not the same as infective endocarditis.

Raw response:
Reasoning:
Unrepaired cyanotic congenital heart disease with right-to-left shunt allows venous thrombi to bypass the lungs and embolize systemically. She is afebrile with normal WBC, only mildly elevated CRP, no peripheral stigmata of endocarditis, and acute CT-proven renal infarct; anticoagulation was initiated for embolism. Infective endocarditis with septic emboli is less likely given absence of fever and classic signs.

¡answer¿
Paradoxical systemic arterial embolism causing acute right renal infarction in cyanotic congenital heart disease (unrepaired tricuspid atresia with d-TGA)

¡/answer¿

MedReason-8B (wrong)

Prediction: Intracardiac thrombus with paradoxical embolism
Judge reason: Predicted diagnosis is intracardiac thrombus with paradoxical embolism/renal infarction, which is not the same as infective endocarditis.

Raw response:
Finding Reasoning Paths:
1. **Right upper quadrant abdominal pain**: Consider causes such as gallbladder disease (e.g., cholecystitis), pancreatitis, peptic ulcer disease, or vascular events like mesenteric ischemia or renal infarction.
2. **History of flu-like illness and doxycycline use**: Evaluate for complications related to recent antibiotic use (e.g., drug-induced side effects) or secondary infections.
3. **Cyanosis and low oxygen saturation**: Investigate cardiac or pulmonary causes, including pulmonary embolism, chronic lung disease, or congenital heart defects.
4. **Acute renal infarction on CT scan**: Explore potential causes such as thromboembolism, vasculitis, or other vascular abnormalities.
5. **Congenital heart disease history**: Assess how underlying structural heart defects (e.g., tricuspid atresia, transposition of the great arteries) contribute to systemic issues like hypoxemia or embolic events.

—

Reasoning Process:
1. **Initial Presentation**:
- The patient has severe right upper quadrant abdominal pain, cyanosis, and hypoxemia (oxygen saturation 79%). These findings suggest a systemic issue involving both cardiovascular and possibly gastrointestinal systems.
- The absence of fever, nausea, vomiting, or diarrhea makes infectious or inflammatory causes less likely in the GI tract.

2. **Cardiac History**:
- The patient has a complex congenital heart defect (tricuspid atresia with transposition of the great arteries). This condition predisposes her to cyanosis due to impaired oxygenation and potential complications like paradoxical emboli.
- Cyanosis and hypoxemia could also result from pulmonary hypertension or right-to-left shunting due to the congenital defect.

3. **Renal Infarction**:
- The CT scan shows an acute right renal infarction, which is often caused by embolism, thrombosis, or vasculitis. In this case, embolism is most likely given the systemic nature of the presentation.
- Emboli can originate from the heart, particularly in patients with congenital heart disease, where intracardiac thrombi may form due to stagnant blood flow or turbulent currents.

4. **Doxycycline Use**:
- Doxycycline is not commonly associated with embolic phenomena, so it is unlikely to be directly responsible for the renal infarction.
- However, the flu-like illness treated with doxycycline raises the possibility of a preceding viral infection that could have triggered a hypercoagulable state or endothelial damage, increasing the risk of thromboembolism.

5. **Systemic Embolism**:
- The combination of cyanosis, hypoxemia, and renal infarction strongly suggests a systemic embolic event. Given the patient's congenital heart disease, paradoxical embolism (via a right-to-left shunt) is a plausible mechanism.
- Intracardiac thrombus formation is a common complication in patients with congenital heart defects, especially when there are atrial or ventricular septal defects allowing shunting of deoxygenated blood.

6. **Exclusion of Other Causes**:
- There is no evidence of infection (no fever, leukocytosis, or elevated CRP).
- No signs of pancreatitis or peptic ulcer disease based on the clinical presentation.
- The absence of neurological symptoms makes stroke or cerebral embolism less likely.

—

Conclusion:
The most likely diagnosis is **intracardiac thrombus formation**, leading to a paradoxical embolism causing the acute right renal infarction and systemic hypoxemia. This conclusion aligns with the patient's congenital heart disease, cyanosis, and renal infarction findings.

Final Answer

Intracardiac thrombus

J AI USE STATEMENT

We used generative AI tools to polish the manuscript for clarity and readability and to assist in finding relevant prior work. LLMs were also used to generate synthetic factual and counterfactual QA data and their reasoning traces. The data generation and validation procedures are detailed in Section 3.2 and Appendix C. The authors reviewed the AI-assisted text and take responsibility for the final content of this work.