MedAction: Towards Active Multi-turn Clinical Diagnostic LLMs

Hsin-Ling Hsu[1] Zizheng Wang[2*] Donghua Zhang[3*] Nai-Chia Chen[1] Jerry Wang[1] Jun-En Ding[4] Chia-Hsuan Hsu[5] Guoan Wang[4] Feng Liu[6] Fang-Ming Hung[7] Chenwei Wu[3†] Liyue Shen[3†]

[1]National Chengchi University [2]Georgetown University [3]University of Michigan [4]Stevens Institute of Technology [5]National Taiwan University of Science and Technology [6]Rutgers University [7]Far Eastern Memorial Hospital

Abstract

Most existing LLM diagnoses are evaluated on static, single-turn settings where complete patient information is provided upfront, an oversimplification of real clinical practice. We study active diagnosis: the real-life clinical process of starting from initial observation, ordering tests, interpreting results, and updating a differential diagnosis across multiple turns. Through systematic analysis, we identify three recurring failure modes in current LLMs — ungrounded test ordering, unreliable diagnostic update, and degraded multi-turn coherence — that together reveal a core deficit: existing medical training data teaches models to reason from complete information but not to act under evolving, partial evidence. To address this gap, we introduce MedAction, a tree-structured distillation pipeline that synthesizes diverse and high-quality multi-turn diagnostic trajectories via LLM–environment interaction. We propose two knowledge-graph-grounded metrics to filter trajectory quality: Disease Trajectory Consistency (DTC), which tracks whether the model's hypothesis converges toward the correct diagnosis, and Reasoning–Action Consistency (RAC), which verifies that belief updates are driven by gathered evidence. Using this pipeline, we construct MedAction-32K, a dataset of 32,681 trajectories from 2,896 PMC cases. Fine-tuning an 8B model on MedAction-32K achieves state-of-the-art performance among open-source models on both MedR-Bench and our curated MedAction-300-Hard benchmark, pushing the edge for open-source medical LLMs.[1]

1 Introduction

Large language models (LLMs) have shown promising performance on medical question answering (QA) benchmarks, suggesting great potential as clinical decision-support tools for diagnosis (Singhal et al., 2023; Jin et al., 2021). However, most existing benchmarks formulate the medical diagnosis as a static, one-shot QA task: the model receives a complete and fully assembled patient record including laboratory results, imaging, and biomarkers, implicitly assuming such data are already available, and then is asked to predict the diagnosis via a well-posed question prompt. This setup is highly simplified and unrealistic and does not reflect the complexities of real-world clinical diagnosis.

Instead, as shown in Figure 1, medical diagnosis is an active, iterative process. In practice, clinicians always begin with incomplete information, then iteratively recommend new tests to resolve uncertainty, and progressively narrow the differential diagnosis based on new test

[*]These authors contributed equally.

[†]Co-senior authorship. Contact: chenweiwu99@gmail.com

[1]Code and data available at https://github.com/JustinHsu1019/MedAction

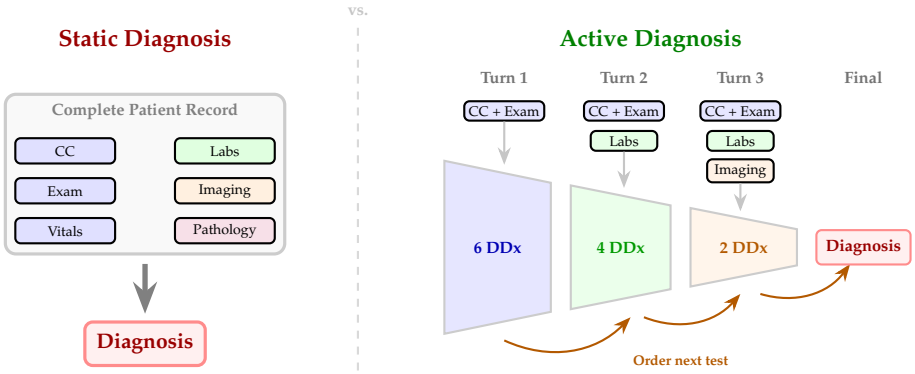

Figure 1: **Static vs. Active Diagnosis.** *Left:* Existing benchmarks provide a complete patient record and ask for a one-shot diagnosis. *Right:* In active diagnosis, the model begins with only the chief complaint (CC) and physical examination, then iteratively orders new tests to resolve uncertainty and progressively narrows the *differential diagnosis (DDx)* from six candidate conditions down to a definitive diagnosis through multiple turns.

results accordingly, until making a final definitive diagnosis with sufficient confidence. We call this new formulation *active diagnosis with iterative test ordering*. To build LLMs that can be genuinely useful in clinical settings, we must enable their capability for active diagnosis via evaluation in such multi-turn, information-gathering regime.

While state-of-the-art LLMs perform well on static medical QA (Jin et al., 2021; Pal et al., 2022; Jin et al., 2019) with growing ability in multi-turn tasks in general domains (Wu et al., 2025b; Liu et al., 2024; Yao et al., 2025), a recent study (Qiu et al., 2025) reveals a striking performance degradation when the same models are evaluated on active medical diagnosis task. This gap persists even for powerful agentic models such as GPT-o3 (OpenAI, 2025) and medical specialist models such as MedGemma (Sellergren et al., 2025). To understand what makes this setting uniquely challenging, we investigate LLM behavior in active multi-turn diagnosis and identify three representative failure modes (Figure 2).

Ungrounded test ordering. The clinical action space spans thousands of laboratory tests, imaging studies, and procedures, resulting in a highly complex decision landscape. We find that LLMs frequently order tests without a clear connection to their stated hypotheses, lacking case-specific clinical reasoning and context-grounded actions.

Unreliable diagnostic update. Interpretation of clinical test results is non-trivial, requiring consideration of competing hypotheses, prior evidence, and patient context. Due to imprecise understanding of newly acquired test results, LLMs often adhere rigidly to an initial diagnosis despite contradictory evidence, or shift erratically between unrelated diagnoses.

Degraded multi-turn coherence. As evidence accumulates across multi-turn interactions, LLMs often lose track of which tests have been ordered, what results were obtained, and what uncertainty remains unresolved, leading to inconsistencies in longer context.

These failures reveal critical limitations of existing medical LLMs, which are usually fine-tuned with medical training data such as chain-of-thought (CoT) reasoning trajectories (Wu et al., 2025a; Chen et al., 2025a): models are trained to *think* — perform reasoning toward a diagnosis assuming complete information — but not to *act* — sequentially decide which new information to gather ("action") and update the current hypothesis ("status") in response to newly collected information. Addressing these challenges in active diagnosis requires the development of multi-turn interactive medical training data that can equip LLMs with such abilities to both reason and act correctly.

In this work, we introduce **MedAction**, a framework for automatically constructing high-quality multi-turn active diagnosis training data from publicly available medical literature. Unlike existing medical CoT data that teaches models to *think* from complete information, MedAction trajectories capture both *thinking* and *acting* under partial, evolving evidence.

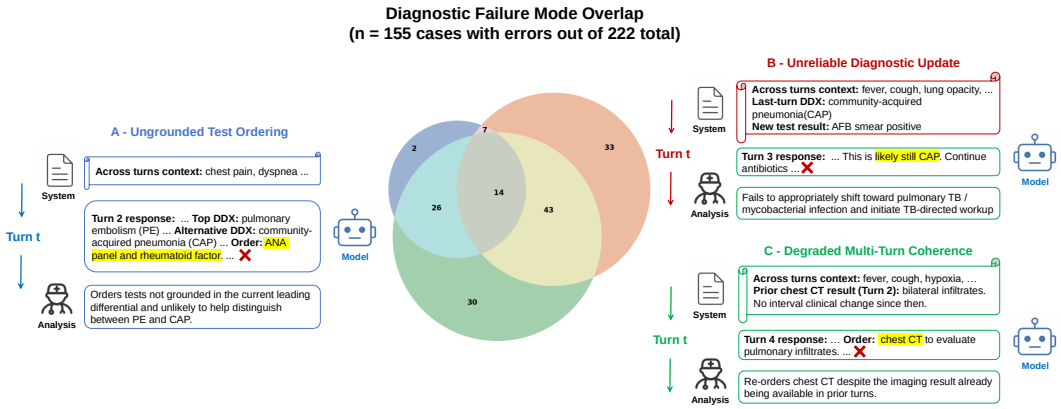

Figure 2: **Three representative failure modes (MedGemma) in active diagnosis.** Each panel illustrates one failure mode with a concrete example: (A) ungrounded test ordering, (B) unreliable diagnostic update, and (C) degraded multi-turn coherence. The Venn diagram shows failure mode co-occurrence across 155 erroneous cases out of 222 examined cases in our training set.

Starting from raw case reports, we first extract 2896 interactive clinical environments that simulate the partial-information setting of active diagnosis: the model observes only the initial patient information and needs to take proactive actions to request new test results turn by turn. We then construct structured multi-turn trajectories by having state-of-the-art LLMs interact with these environments, producing chains of reasoning-action data that mirror real clinicians' diagnostic procedure: maintaining a differential, identifying diagnostic uncertainty, and selecting tests to resolve it.

To distill high-quality training data and maximize data utilization, we propose two knowledge-graph-grounded metrics to enable clinically grounded rejection sampling. **Disease Trajectory Consistency (DTC)** measures the knowledge-graph distance between the model's current top hypothesis and the ground-truth diagnosis on PrimeKG (Chandak et al., 2023), allowing us to retain complete trajectories that reach the correct answer as well as partially correct trajectories where the hypothesis was still converging. **Reasoning–Action Consistency (RAC)** then measures whether ordered tests drove meaningful changes in the differential on ClinGraph (Johnson et al., 2024), filtering individual turns where hypothesis shifts were unrelated to gathered evidence. We further investigate key design choices in trajectory construction, including teacher model selection, tree sampling strategy, and filtering thresholds, providing actionable guidance for building multi-turn medical reasoning-action data. Training on the resulting **MedAction-32K dataset**, our 8B-parameter model achieves state-of-the-art performance among open-source models under multi-turn active diagnosis settings, demonstrating that high-quality trajectory data, even at modest scale, drives larger gains than architecture scaling alone. In addition to the training set, we curate **MedAction-300 Hard**, a test set of 300 rare-disease cases that demand longer diagnostic trajectories and more extensive test ordering than existing benchmarks.

Our contributions are threefold:

1. **New formulation and failure analysis.** We reformulate the clinical diagnosis as multi-turn active diagnostic setting with iterative test ordering, and identify three representative failure modes that explain why current LLMs struggle in this setting.

2. **New dataset construction and new metrics.** We introduce a tree-structured data distillation pipeline that automatically synthesizes high-quality multi-turn medical diagnostic trajectories from public case reports, investigate key design choices in trajectory construction, and release the MedAction-32K training set and the MedAction-300 Hard test set—a challenging benchmark of rare-disease cases—to the community. We also propose two new metrics for training-data distillation and rigorous model evaluation.

Feature	Static Benchmarks (Jin et al., 2021) (Jin et al., 2019)	Interactive Inquiry (Li et al., 2024) (Lai et al., 2026)	Multi-Agent Simulation (Schmidgall et al., 2024)	Multi-turn Eval (Qiu et al., 2025)	Training Data Construction (Wu et al., 2025a) (Ossowski et al., 2025)	MedAction (Ours)
Multi-turn diagnosis	✗	✓	✓	✓	✗	✓
Iterative test ordering	✗	✗	✓	✓	✗	✓
Interactive data generation	✗	✗	✗	✗	✗	✓
Turn-level quality filtering	✗	✗	✗	✗	✗	✓
Data recipe	✗	✓	✗	✗	✓	✓
Evaluation benchmark	✓	✓	✓	✓	✗	✓

Table 1: Comparison with existing medical benchmarks and training frameworks.

3. **Efficient and effective model.** Our dataset trains an 8B-parameter model that achieves state-of-the-art results amongst open-source models on active diagnosis and test recommendation.

2 Related Work

We summarize the comparison with prior work in Table 1.

Beyond Static Diagnosis LLM Benchmarks Standard medical benchmarks such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019) evaluate LLMs in a static, single-turn format with complete records. Recent work has begun to study more realistic settings: MediQ (Li et al., 2024) converts static MedQA cases into partial-information scenarios requiring models to ask follow-up questions; MedR-Bench (Qiu et al., 2025) evaluates LLMs on test recommendation and diagnosis over multiple turns, finding that accuracy drops sharply on examination recommendation even for models performing well on static tasks. Other efforts include MedJourney (Wu et al., 2024), which decomposes the clinical process into sequential stages; AMIE (Tu et al., 2025), which studies conversational diagnostic AI; and Liu et al. (Liu et al., 2025), which analyze challenges in multi-turn diagnostic LLM systems. Despite this progress, the underlying causes of failure in active diagnosis, and how to construct training data for this setting, remain unexplored.

Medical LLM Training Data Construction Carefully curated reasoning traces (Muennighoff et al., 2025; Guha et al., 2026) can significantly improve reasoning performance. Applying these ideas to medicine requires reasoning chains to be factually accurate. MedReason (Wu et al., 2025a) tackles this by grounding CoT generation in a medical knowledge graph; while other works distill long reasoning traces from stronger models with domain-specific verification (Ossowski et al., 2025; Chen et al., 2025a; Huang et al., 2025). UltraMedical (Zhang et al., 2024) provides a large-scale biomedical training resource. However, all existing medical data construction work targets single-turn reasoning from complete information, neglecting the real-life active diagnosis setup. MedAction directly addresses this gap by constructing trajectories that capture both the reasoning and actions at each diagnostic step.

3 Methodology

3.1 Problem Formulation

We define **active diagnosis with iterative test ordering** as a sequential decision-making problem under partial observability. Let s^* denote the ground-truth diagnosis. A diagnostic trajectory $\tau = (o_1, a_1, r_1, o_2, a_2, r_2, \ldots, o_T)$ consists of: **observations** o_t, the accumulated clinical evidence at turn t (initial presentation plus all prior test results), where partial observability arises because only results of ordered tests are revealed; **actions** $a_t \subseteq \mathcal{A}$, clinical tests selected from a large open-vocabulary action space; **results** r_t, the findings returned for a_t (lab values, imaging reports, or clinical observations); and **belief state** b_t, the model's ranked differential diagnosis approximating the posterior over s^*. The process terminates when the model commits to a definitive diagnosis or the turn limit $T_{\max}$ is reached.

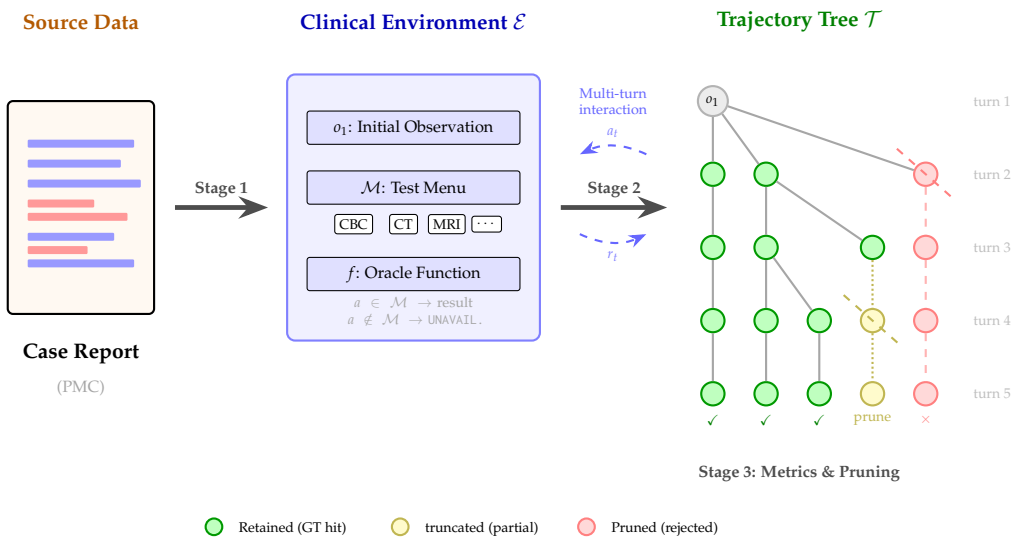

Figure 3: **Overview of the MedAction pipeline.** Stage 1 transforms retrospective case reports into interactive clinical environments. Stage 2 generates a trajectory tree through multi-turn LLM–environment interaction. Stage 3 applies knowledge-graph-grounded metrics to retain high-quality trajectories: complete trajectories reaching the ground-truth diagnosis (green, ✓), truncated prefixes from partially correct paths (yellow), while rejecting divergent or ungrounded trajectories (red, ×).

3.2 Overview

MedAction addresses the scarcity of high-quality multi-turn diagnostic training data through a three-stage pipeline (Figure 3):

$$\underbrace{\text{Case Reports}}_{\text{PMC}} \xrightarrow{\text{Stage 1}} \underbrace{\text{Clinical Environment}}_{\mathcal{E}} \xrightarrow{\text{Stage 2}} \underbrace{\text{Trajectory Tree}}_{\mathcal{T}} \xrightarrow{\text{Stage 3}} \underbrace{\text{Training Data}}_{\mathcal{D}} \tag{1}$$

We first transform each case report into an interactive *clinical environment* (Stage 1), generate tree-like diagnostic trajectories by having state-of-the-art LLMs interact with this environment (Stage 2), and filter the resulting trajectories using knowledge-graph-grounded metrics (Stage 3).

3.3 Data Generation Framework

Stage 1: Clinical Environment Construction From each PMC-OA case report (National Library of Medicine, 2003), we use an LLM to extract a clinical environment $\mathcal{E} = (o_1, \mathcal{M}, f)$: the **initial observation** o_1 (chief complaint, history, physical exam); the **test menu** $\mathcal{M} \subseteq \mathcal{A}$ of documented tests; and an **oracle function** $f : \mathcal{A} \rightarrow \mathcal{R} \cup \{\texttt{UNAVAILABLE}\}$ returning documented results or UNAVAILABLE.

Stage 2: Multi-turn Chain-of-Reasoning-Action Trajectory Distillation Rather than learning a black-box policy $\pi(a_t \mid o_t)$, we require the model to follow an explicit reason-action pattern at each turn:

$$o_t \xrightarrow{\text{analyze}} b_t \xrightarrow{\text{identify}} p_t \xrightarrow{\text{select}} a_t \tag{2}$$

where b_t is a ranked differential with confidence estimates, p_t states the key remaining uncertainty distinguishing top hypotheses, and a_t selects tests justified by p_t. The oracle returns $r_t = f(a_t)$, appended to o_t for the next turn. To prevent context degradation, we maintain a cumulative record of all tests and results alongside the model's reasoning from the last two turns.

To maximize diversity, we employ tree-structured sampling: We sample K independent rollouts from o_1 using multiple teacher LLMs, as well as branched rollouts launched from intermediate states of existing trajectories, producing diverse continuations from shared evidence states.

Stage 3: Metrics and Data Pruning Not all sampled trajectories exhibit coherent diagnostic reasoning. We define two knowledge-graph-grounded metrics to filter them, derived from shortest-path distances on PrimeKG (Chandak et al., 2023) (disease hypothesis–disease ground truth relationships) and ClinGraph (Johnson et al., 2024) (test–disease relationships), with free-text entities mapped to graph nodes via a biomedical information retrieval model (Jin et al., 2023).

Disease Trajectory Consistency (DTC). DTC measures the distance between the model's current top hypothesis and the ground-truth diagnosis s^*, quantifying whether the diagnostic trajectory is converging toward or diverging from the correct answer. Using PrimeKG:

$$\mathrm{DTC}_t = h_{\mathrm{PrimeKG}}(b_t^{(1)},\ s^*) \tag{3}$$

where $b_t^{(1)}$ denotes the top-ranked diagnosis in the differential diagnosis list b_t at turn t, and h_{PrimeKG} denotes the shortest-path distance on PrimeKG. A trajectory with decreasing DTC over turns indicates convergence toward the correct diagnosis; increasing DTC signals divergence. $\mathrm{DTC}_t = 0$ means the model has identified the correct diagnosis at turn t. For pruning, we scan backward from the final turn T, searching for the latest turn t^* satisfying $\mathrm{DTC}_{t^*} \leq \mathrm{DTC}_1$, that is, the latest turn at which the model's hypothesis is at least as close to the ground truth as its initial guess. All turns after t^* are truncated. If no such turn $t^* \geq 2$ exists (i.e., every subsequent turn is strictly farther from the ground truth than the initial hypothesis), the trajectory is discarded entirely, as it demonstrates no diagnostic convergence despite multiple rounds of evidence gathering.

Reasoning–Action Consistency (RAC). RAC measures whether the tests ordered in the previous turn actually drove meaningful changes in the differential diagnosis, providing a link between actions and belief updates. Let $\Delta b_t = (b_t \setminus b_{t-1}) \cup (b_{t-1} \setminus b_t)$ denote the set of diagnoses that were added or removed between turns $t-1$ and t, identified by matching free-text diagnoses to knowledge graph node identifiers rather than surface string comparison. For each changed diagnosis $d \in \Delta b_t$, we compute its minimum shortest-path distance to any action ordered in the previous turn:

$$\mathrm{RAC}_t = \frac{1}{|\Delta b_t|} \sum_{d \in \Delta b_t} \min_{a' \in a_{t-1}} h_{\mathrm{ClinGraph}}(d,\ a'), \quad t \geq 2 \tag{4}$$

Lower RAC indicates evidence-driven updates; higher values indicate arbitrary hypothesis shifting. For each turn $t \geq 2$ retained after DTC pruning, if $\mathrm{RAC}_t > \tau_{\mathrm{RAC}}$ (a maximum hop threshold), we remove the preceding turn $t-1$ that ordered ineffective tests. This pruning is applied in a single pass.

4 Experiments

We train an II-Medical-8B (Internet, 2025) model based on Qwen-3 8B (Yang et al., 2025) model using supervised fine-tuning (SFT) on MedAction-32K, which consists of 32681 samples from 2896 cases distilled using Baichuan-M3 (Dou et al., 2026) and GPT5.4-mini (OpenAI, 2026b) as our teachers. We evaluate on the Med-R (Qiu et al., 2025) dataset and our own MedAction-300 Hard test dataset (Appendix B). We compare our model against open-source general domain models including Qwen-QwQ (32B) (Team, 2025; Yang et al., 2024), closed-source models including OpenAI-o3-mini (OpenAI, 2025), Gemini 2.5 Flash-Lite (Comanici et al., 2025), and GPT-5.4 (OpenAI, 2026a), and medical domain open-source models including Baichuan-M1 (14B) (Wang et al., 2025), Baichuan-M3 (235B-8bit) (Dou et al., 2026), MedGemma (27B) (Sellergren et al., 2025), and Diagnosis-GPT (34B) (Chen et al., 2025b). We report average test recommendation precision, recall and F1, as well as multi-turn diagnosis

Model	MedR-Bench				MedAction-300-Hard			
	Prec	Rec	F1	Diag Acc	Prec	Rec	F1	Diag Acc
Closed-Source Models								
OpenAI-o3-mini	0.36	0.42	0.39	0.71	**0.54**	0.47	**0.50**	0.67
Gemini 2.5 Flash-Lite	0.36	0.44	0.40	0.68	0.47	0.48	0.47	0.63
GPT-5.4	**0.47**	**0.56**	**0.51**	**0.81**	0.42	**0.60**	0.49	**0.82**
Open-Source Models								
Baichuan-M1 (14B)	0.42	0.39	0.40	0.61	0.47	0.41	0.44	0.56
Baichuan-M3 (235B)	0.52	0.52	0.52	0.69	0.50	**0.54**	**0.52**	0.66
MedGemma (27B)	0.39	0.46	0.42	0.65	**0.52**	0.52	**0.52**	0.63
Qwen-QwQ (32B)	0.35	0.49	0.41	0.66	0.51	0.41	0.45	0.58
II-Medical-8B	**0.59**	0.36	0.45	0.51	0.45	0.45	0.45	0.46
II-Medical-8B SFT (Ours)	0.54	**0.58**	**0.56**	**0.74**	0.51	0.51	0.51	**0.69**

Table 2: Results on MedR-Bench and MedAction-300-Hard.

accuracy across 3 runs. In addition, we conduct a blinded expert evaluation on a subset of MedAction-300-Hard to assess the clinical quality of diagnostic trajectories and compare expert judgments with GPT-4o evaluation. More training and inference configs are included in Appendix C. Prompt templates are provided in Appendix A.

5 Results

5.1 Main Results

As shown in Table 2, among open-source models, we achieve the highest diagnosis accuracy, outperforming one of our teachers Baichuan-M3 (235B). We also achieve consistently high test recommendation precision and recall, which demonstrates that high-quality multi-turn trajectory data drives larger gains than model scale alone.

GPT-5.4 is the strongest diagnosis model, establishing a ceiling with meaningful room for improvement. The base II-Medical-8B achieves high precision but very low recall, reflecting the ungrounded test ordering failure mode (Figure 2); MedAction training alleviates this issue substantially.

5.2 Human Expert Validation

To complement automatic evaluation, we conducted a blinded expert study on 100 randomly selected MedAction-300-Hard cases, resulting in 300 model trajectories (Base II-Medical-8B, OpenAI-o3-mini, and MedAction-SFT). An experienced licensed clinician independently evaluated each trajectory without access to the model identity or GPT-4o assessment. The evaluation covered diagnostic correctness, test appropriateness, safety, redundancy, and sequencing quality.

As shown in Table 3, MedAction-SFT achieved the strongest overall clinical performance, substantially improving appropriate test ordering while reducing redundant examinations without sacrificing safety. These findings are consistent with the automatic evaluation results. Furthermore, GPT-4o showed strong agreement with the blinded clinician (Cohen's $\kappa = 0.81$, 91% raw agreement) while preserving the same ranking across all models, suggesting that GPT-4o provides a reliable approximation of human assessment.

5.3 Analysis

A central question in MedAction is: *how should we curate multi-turn diagnostic trajectories to teach LLM proper active diagnosis?* Hence, we explain via experiments four design choices in the pipeline. All results in the analysis section are on MedR-Bench.

Metric	II-Medical	o3-mini	MedAction-SFT
Human diagnosis accuracy	0.44	0.59	**0.66**
Appropriate tests ↑	39.9	48.3	**66.5**
Safe tests ↑	**89.6**	86.1	89.0
Redundant tests ↓	49.1	35.8	**18.4**
Good sequencing ↑	22	36	**49**
GPT-4o diagnosis accuracy	0.48	0.62	**0.70**

Table 3: Blinded clinician evaluation. GPT-4o achieved Cohen's $\kappa = 0.81$ (91% agreement) with the clinician.

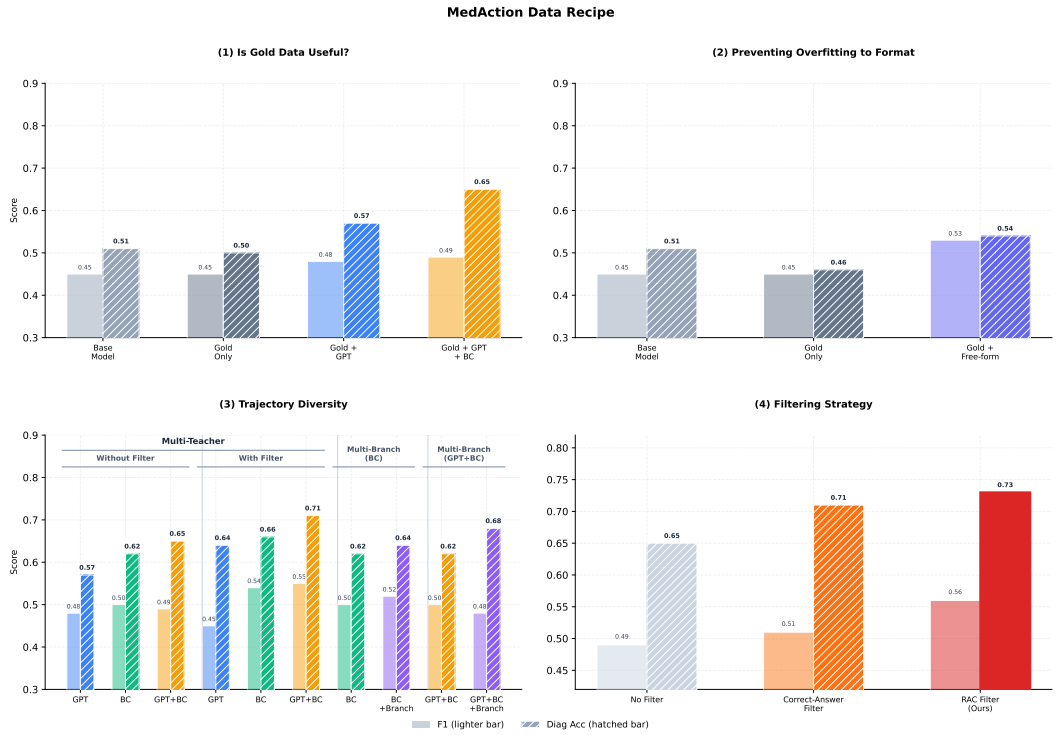

Figure 4: Ablation studies on trajectory data curation. Lighter bars: test recommendation F1; hatched bars: diagnosis accuracy.

RQ1: Can "Gold" trajectories extracted directly from case reports serve as effective training data? An intuitive approach to constructing multi-turn training data is to extract established diagnostic trajectories from public medical case reports. We call these *gold trajectories*: a strong LLM reconstructs the documented sequence of tests and findings from the raw case text, producing trajectories that are by construction ideal—every ordered test returns a result, every result is informative, and the trajectory converges to the correct diagnosis without detours. An alternative is to *distill trajectories*: rather than reconstructing a documented path, we have teacher LLMs (GPT-5.4-mini and Baichuan-M3) interact with the clinical environment under genuine partial observability, where tests can be unavailable, results may be inconclusive, and diagnostic repivots are necessary. We compare four configurations: the base model (no SFT), gold-only, gold + GPT-distilled, and gold + GPT + Baichuan-distilled.

As shown in Figure 4-(1), training on gold trajectories alone does not improve diagnosis accuracy or test recommendation F1 over simply prompting the base model to perform structured diagnosis. We inspected model outputs and found out that the model learns the reasoning format of active differential diagnosis well but easily gets stuck when the environment deviates from a perfect path.

Takeaway: Gold trajectories teach output format but produce weak reasoning. Environment-interactive distillation sampling is essential.

RQ2: Does the output schema affect reasoning flexibility? RQ1 raises concerns about the model only learning the format but not the way to reason and act. All trajectories in our dataset follow a fixed output schema at each turn: differential diagnosis list, pivot statement, and action selection. While this structure aids mimicking how clinicians reason, we found that the model's "thinking mode" behaviors, such as self-correction, reflection, and exploratory hypothesizing are slightly reduced. We address this by adding a small proportion (10%) of *free-form trajectories*: sequences generated without a prescribed output structure, where the model reasons unconstrained before committing to an action. As shown in Figure 4-(2), mixing in free-form data improves both test recommendation F1 and diagnosis accuracy over the gold-only baseline. The gain is most pronounced in F1, suggesting that free-form reasoning restores the model's ability to consider a broader set of tests rather than defaulting to schema-conditioned habits.

Takeaway: A small proportion of free-form data mitigates format overfitting and rescues thinking-mode behaviors, yielding meaningful gains in test recommendation quality.

RQ3: What sources of diversity matter most in rejection sampling? Diversity in training trajectories is particularly important for active diagnosis because the task is inherently *multi-path*: unlike static QA where a single reasoning chain leads to the answer, the same initial presentation can support multiple valid diagnostic strategies—a clinician may screen broadly with routine labs before narrowing, or pursue a targeted confirmatory test immediately. Moreover, the action space is vast and open-vocabulary, meaning any single teacher LLM will cover only a narrow slice of valid test-ordering behavior. We examine two orthogonal sources of diversity that address different aspects of this challenge.

Multi-teacher diversity. Different teacher LLMs exhibit different clinical reasoning styles—for instance, one may favor broad screening while another pursues targeted confirmatory tests. We compare trajectories generated by GPT-5.4-mini alone, Baichuan-M3 alone, and the combination, under two conditions: without any quality filtering, and with correctness filtering (retaining only trajectories that reach the ground-truth diagnosis). *Multi-teacher diversity* captures variation in clinical reasoning *style*. Without filtering, combining teachers yields moderate gains over single-teacher baselines. With filtering, the same combination produces a substantially larger improvement, showing that multi-teacher diversity is most valuable *after* low-quality trajectories are removed: unfiltered mixing introduces noise that dilutes the benefit of stylistic diversity.

Multi-branch diversity captures variation in diagnostic *decisions* from identical evidence states. We generate branched trajectories by launching new rollouts from intermediate turns of existing trajectories, producing multiple valid continuations from the same accumulated evidence. This directly mirrors the branching nature of clinical reasoning: given the same labs, two clinicians may legitimately order different next tests. Branching consistently improves performance in both single-teacher and multi-teacher configurations.

Takeaway: Multi-teacher diversity broadens reasoning style coverage and requires filtering to realize max benefits; multi-branch diversity also provides consistent gains.

RQ4: Does knowledge-graph-grounded filtering outperform correctness-only filtering? From RQ3 analysis we identified filtering as an effective approach to generate high quality data and provide performance gains. We now compare *what* to filter on. We evaluate three strategies applied to the same pool of GPT+BC trajectories with branching: (1) *no filtering*, which retains all sampled trajectories regardless of outcome; (2) *correctness filtering*, which retains only complete trajectories that reach the ground-truth diagnosis; and (3) *DTC+RAC filtering* (ours), which additionally retains convergent partially correct trajectories and removes turns where diagnostic belief changes were ungrounded in the evidence gathered (Section 3). As shown in Figure 4-(4), unfiltered multi-teacher trajectories already achieve

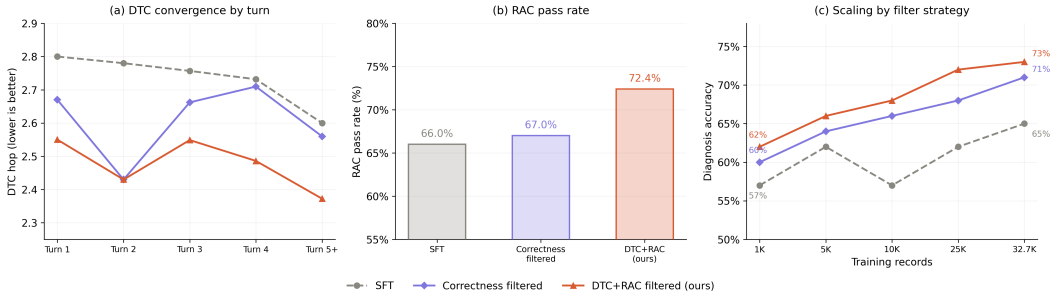

Figure 5: **Effect of trajectory filtering on diagnostic performance.** ***(a)*** Average DTC for differently finetuned models at each turn. Lower values indicate closer alignment. ***(b)*** Turn level pass rate under the RAC filtering. Higher is better. ***(c)*** Scaling performance for each filtering strategy.

reasonable performance, but both filtering strategies yield meaningful additional gains. Correctness filtering improves diagnosis accuracy substantially but leaves F1 relatively low, because it retains only trajectories with correct endpoints while discarding all partially correct reasoning. Our filtering improves on both metrics, with the largest gain in F1. The mechanism is straightforward: by retaining convergent prefixes and pruning only ungrounded turns, RAC filtering recovers useful training signal that correctness filtering discards—particularly the test-ordering and belief-updating behavior in early and middle turns of otherwise imperfect trajectories.

To understand *how* this translates into improved model behavior, we evaluate 100 held-out cases across the three training configurations and measure trajectory-level diagnostic quality (Figure 5). Panel (a) tracks the DTC hop distance between each model's top hypothesis and the ground-truth diagnosis at each turn. The SFT baseline starts with the highest DTC and converges slowly, while correctness-filtered and DTC+RAC-filtered models both start lower. The DTC+RAC model reaches the lowest final DTC, confirming that it learns effective diagnostic reasoning over longer horizons rather than committing early to a wrong answer. Panel (b) applies RAC filtering to all model outputs, and the pass rate measures the fraction of turns where belief updates were grounded in the evidence gathered. The DTC+RAC-filtered model achieves the highest pass rate, directly addressing the reasoning-action mismatch. Panel (c) examines data efficiency: we train models at five data scales (1K–32K records) under all three filtering strategies. DTC+RAC filtering dominates at every scale, with the largest advantage at low data budgets. Unfiltered scaling is non-monotonic, dipping at 10K before recovering, while DTC+RAC scaling increases monotonically, confirming that filtering removes harmful training signal rather than merely selecting easier cases.

Takeaway: DTC+RAC filtering outperforms correctness-only filtering by recovering useful intermediate reasoning from partially correct trajectories and reducing ungrounded turn-level actions. Models trained on DTC+RAC-filtered data exhibit stronger diagnostic convergence, reduction in ungrounded actions, and more data-efficient scaling.

6 Conclusion and Limitations

We presented MedAction, a data construction framework for improving LLM performance on active diagnosis. Our analysis identifies three systematic failure modes in current LLMs, and our environment-interactive trajectory distillation — grounded by two new metrics (DTC and RAC) — directly addresses them. Our 8B model fine-tuned on MedAction-32K achieves state-of-the-art open-source performance, providing strong foundations for future work. A natural next step is extending our framework to post-training approaches such as reinforcement learning, for which MedAction's environments and metrics provide direct building blocks. We also plan to conduct more clinical validation and broader test coverage to ensure real-world applicability.

References

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023. URL https://doi.org/10.1038/s41597-023-01960-3.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. Towards medical complex reasoning with LLMs through medical verifiable problems. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14552–14573, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.751. URL https://aclanthology.org/2025.findings-acl.751/.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. CoD, towards an interpretable medical agent using chain of diagnosis. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14345–14368, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.740. URL https://aclanthology.org/2025.findings-acl.740/.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Chengfeng Dou, Fan Yang, Fei Li, Jiyuan Jia, Qiang Ju, Shuai Wang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Hongda Zhang, et al. Baichuan-m3: Modeling clinical inquiry for reliable medical decision-making. *arXiv preprint arXiv:2602.06570*, 2026.

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Rea Sprague, Ashima Suvarna, Benjamin Feuer, Leon Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah M Pratt, Vivek Ramanujan, Jon Saad-Falcon, Stutee Acharya, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alex Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7xjoTuaNmN.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *arXiv preprint arXiv:2504.00869*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Intelligent Internet. Ii-medical-8b: Medical reasoning model, 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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Ruth Johnson, Uri Gottlieb, Galit Shaham, Lihi Eisen, Jacob Waxman, Stav Devons-Sberro, Curtis R. Ginder, Peter Hong, Raheel Sayeed, Ben Y. Reis, Ran D. Balicer, Noa Dagan, and Marinka Zitnik. Clinvec: Unified embeddings of clinical codes enable knowledge-grounded ai in medicine. *medrxiv*, 2024. URL https://www.medrxiv.org/content/10.1101/2024.12.03.24318322.

Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vQGHTyL0Jw.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.

Benjamin Liu, Sejong Kim, Drona Thoka, Varun Puttagunta, Kaylin Sheng, Mark Li, Kiran Nijjer, Adnan Ahmed, Thi Uyen Hanh Le, Sai Chidvilas Gudiboina, Ali Ugur, and Kevin Zhu. Multi-turn LLM systems for diagnostic decision-making: Considerations, biases, and challenges. In *First Workshop on Multi-Turn Interactions in Large Language Models*, 2025. URL https://openreview.net/forum?id=It7AgR3A9H.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

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.

National Library of Medicine. Pmc open access subset [internet]. https://pmc.ncbi.nlm.nih.gov/tools/openftlist/, 2003. Bethesda (MD).

OpenAI. Openai o3-mini system card. Technical report, OpenAI, 2025. URL https://cdn.openai.com/o3-mini-system-card-feb10.pdf.

OpenAI. GPT-5.4 Thinking System Card. Technical report, OpenAI, March 2026a. URL https://openai.com/index/gpt-5-4-thinking-system-card/.

OpenAI. Introducing gpt-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/, 2026b.

Timothy Ossowski, Sheng Zhang, Qianchu Liu, Guanghui Qin, Reuben Tan, Tristan Naumann, Junjie Hu, and Hoifung Poon. Octomed: Data recipes for state-of-the-art multimodal medical reasoning. *arXiv preprint arXiv:2511.23269*, 2025.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, et al. Quantifying the reasoning abilities of llms on clinical cases. *Nature Communications*, 16(1):9799, 2025.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

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.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025.

Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, et al. Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671*, 2025.

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.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. CollabLLM: From passive responders to active collaborators. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=DmH4HHVb3y.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Zhenxi Lin, et al. Medjourney: Benchmark and evaluation of large language models over patient clinical journey. *Advances in Neural Information Processing Systems*, 37:87621–87646, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R Narasimhan. {τ}-bench: A benchmark for \underline{T}ool-\underline{A}gent-\underline{U}ser interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081, 2024.

A Prompt Templates

System prompt for the diagnostic agent.

```
You are an experienced clinical physician performing a structured diagnostic workup.
```

Initial-turn prompt for structured diagnostic reasoning.

```
You are seeing this patient for the first time. Analyze the case and respond using
    EXACTLY the section headers below in EXACTLY this order. Do not skip or rename
    any section.

Patient Case Summary:
{case}

---

### Chain of Thought:
[Think step by step. Label each step <step 1>, <step 2>, etc.]

### DDx List:
[Ranked differentials. Be specific.]
1. Diagnosis name - one sentence reasoning
2. Diagnosis name - one sentence reasoning
3. Diagnosis name - one sentence reasoning

### Pivot:
[One paragraph: the single most important diagnostic question right now and why.]

### Primary Actions:
[Most targeted, high-yield tests to resolve top differentials THIS turn. Specific
    test names only.]
1. test name - purpose
2. test name - purpose

### Additional Information Required:
[Any broader workup beyond Primary Actions: routine labs, imaging, history.
If a test comes back normal or shows no significant findings, do NOT order
any further tests in the same category - move to a completely different
diagnostic category or test a different aspect of the presentation.
Write "Not required." only if you are fully confident in the final diagnosis.]
1. category: specific test or data needed

### Diagnostic Status:
[Write EXACTLY one word: either DONE or CONTINUE.
Write DONE only if you have sufficient information to make a definitive final
    diagnosis right now.
Write CONTINUE if you need more test results before concluding.]

### Conclusion:
[Your single best current diagnosis. Be specific. One sentence only.]
```

Follow-up-turn prompt with cumulative test history and new results.

```
Patient Case Summary:
{case}

--- GLOBAL STATUS (cumulative across all turns) ---

All Tests Ordered So Far:
{done_tests_block}

Tests Confirmed UNAVAILABLE (do NOT re-order these):
{unavailable_tests_block}
NOTE: When key diagnostic tests are UNAVAILABLE, this is a signal that you cannot
    confirm your current leading hypothesis through the expected pathway. You
    should treat UNAVAILABLE tests as a reason to PIVOT your diagnostic direction
    --- consider alternative diagnoses that can be confirmed or excluded using
    different, available tests rather than persisting with the same differential.

--- RECENT DIAGNOSTIC HISTORY (last {window_size} turn(s)) ---

{recent_turns_block}

--- NEW RESULTS THIS TURN ---

{oracle_results}

---

Respond using EXACTLY the section headers below in EXACTLY this order.

### Chain of Thought:
[Step-by-step reasoning.
CRITICAL RULES:
- If key tests are UNAVAILABLE, do NOT simply wait or repeat the same differential.]

### DDx List:
[Updated differentials. Reflect that Normal results have excluded certain
    possibilities and that UNAVAILABLE tests may mean your leading hypothesis
    cannot be verified --- consider alternative diagnoses.]

### Pivot:
[Explain how Normal results and/or UNAVAILABLE tests shifted your diagnostic
    direction and what different path you are pursuing now.]

### Primary Actions:
[New targeted tests ONLY.
RULES:
1. If a test is listed under "Tests Confirmed UNAVAILABLE" above, you are FORBIDDEN
    from re-ordering it.
2. If a test is listed under "All Tests Ordered So Far", do NOT re-order it.
3. You must either:
   a) Order a different, alternative diagnostic test that IS available.
   b) Move to a final "Conclusion" if no other available tests can help.
4. DO NOT use words like "Await", "Wait for", or "Review previous tests".]

### Additional Information Required:
[Any broader workup. Write "Not required." if you are shifting to a final conclusion
    .]

### Diagnostic Status:
[Write EXACTLY one word: either DONE or CONTINUE.
Write DONE only if you have sufficient information right now to make a definitive
    final diagnosis.
```

```
Write CONTINUE if further workup is still needed.]

### Conclusion:
[Single best current diagnosis given all available info.]
```

Oracle prompt for returning ancillary test findings

```
You are a hospital administrator providing status updates on diagnostic requests.

Guidelines:
1. If the requested test is present in the Ancillary Test Results below, provide the
      specific findings.
2. If the test is NOT in the data, respond with:
   "[test name]: This test is currently UNAVAILABLE due to equipment maintenance or
     lack of specialized personnel. You must proceed with clinical diagnosis or
     alternative available testing."

Patient Case
{full_case_summary}

Ancillary Test Results
{anc}
```

Prompt for extracting ordered tests into a JSON array.

```
You are a medical text organizer.
Given a physician's full diagnostic response, extract ALL diagnostic tests or
    examinations they have requested or ordered into a flat JSON array.

Rules:
- Extract ONLY test/examination names, not purposes or reasoning.
- Split compound entries: "CBC, CMP, LFTs" into three separate items.
- Look in all sections: Primary Actions, Additional Information Required, or
    anywhere tests are mentioned.
- If no tests are requested, return [].
- Be specific: "serum troponin I" not just "troponin".

Return ONLY a valid JSON array, no markdown, no extra text:
["test name 1", "test name 2", ...]
```

Prompt for matching predicted tests against ground-truth tests.

```
You are a medical test matching expert.

Given PREDICTED tests and GROUND TRUTH (GT) tests, determine which GT items are
    covered.

RULES:
1. Name equivalence: "CBC" == "Complete Blood Count", "CXR" == "Chest X-ray"
2. Parent covers children: "MRI Brain" covers "MRI Brain T1", "MRI Brain FLAIR"
3. Compound GT items (comma/slash separated): treat as ONE unit.
4. One predicted test can cover multiple GT items. One GT item counted once only.
5. Do NOT mark covered if overlap is minor.

Return ONLY valid JSON:
{
  "gt_covered":   ["..."],
  "gt_uncovered": ["..."],
  "pred_used":    ["..."],
  "pred_unused":  ["..."]
}
```

Prompt for evaluating diagnostic equivalence.

```
You are a medical evaluation assistant.
Determine if Predicted Diagnosis and Ground Truth Diagnosis are clinically
    equivalent.
Return ONLY valid JSON: {"match": true, "reasoning": "..."}
```

Figure 6: Dataset composition across primary body systems and disease types. The outer ring shows the distribution by primary body system, while the inner ring shows the distribution by disease type. The two rings represent independent category views and do not imply a hierarchical relationship.

B Dataset Composition and Test Distribution

Figure 6 shows the distribution of the 2,896 source PMC case reports by primary body system and disease type. Figure 7 summarizes the distribution of diagnostic test categories across these cases.

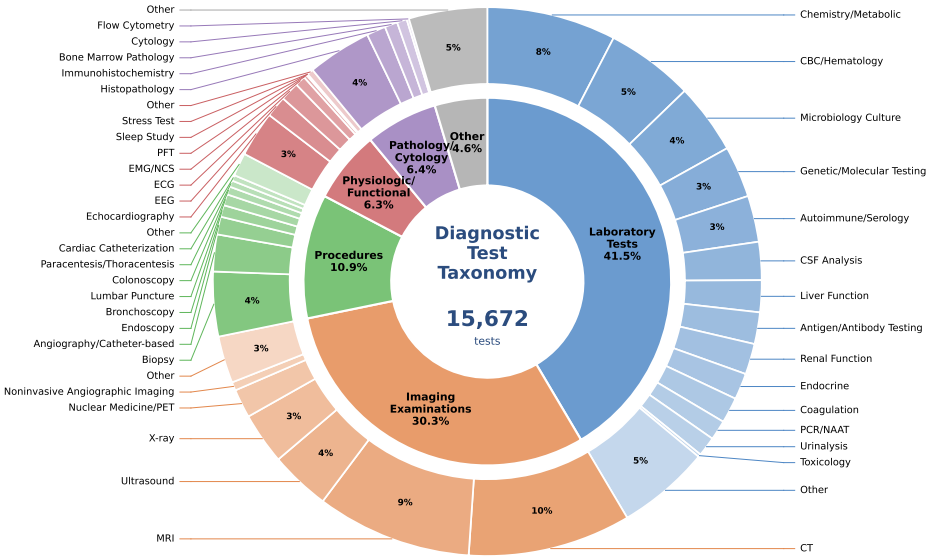

Figure 7: Distribution of diagnostic tests in the dataset. The inner ring shows major test categories, and the outer ring shows representative minor categories within each major category. The center number (15,672) denotes the total number of diagnostic tests in the dataset. Laboratory tests and imaging examinations account for the largest shares of the test space, followed by procedures/invasive diagnostic tests, physiologic/functional tests, and pathology/cytology.

MedAction-300 Hard test dataset. MedAction-300-Hard comprises 300 cases specifically selected for rare and uncommon diseases that are distinct from the other evaluation dataset, MedR-Bench. These cases are deliberately chosen because rare diseases pose a greater diagnostic challenge: they are underrepresented in training corpora, often mimic more common conditions, and typically require longer and more targeted test-ordering sequences to reach a definitive diagnosis. During construction, we used GPT-4o as a validator to ensure no data overlap between MedAction-32K, MedAction-300-Hard and MedR-Bench. We evaluate our model on both the MedAction-300 Hard test set and the MedR-Bench test set. Compared with MedR-Bench, MedAction-300-Hard poses a substantially more challenging diagnostic setting. Cases in our benchmark require more diagnostic actions at each step, as shown by a higher average number of ordered tests per turn (1.92 vs. 1.57), and unfold over longer diagnostic trajectories, with a higher average number of turns per case (6.81 vs. 4.50). Together, these differences suggest that MedAction-300-Hard demands more extensive evidence acquisition and more sustained multi-turn clinical reasoning, making it a more rigorous benchmark for evaluating active diagnostic decision-making.

C Training and Inference Configs

The max turns are set to 8. GPT-4o (Hurst et al., 2024) is used as the evaluator, to be consistent with original Med-R benchmark. We used a learning rate of 2e-4, 3 epochs, and a batch size of 32. Temperature was set to 0.6 and max output tokens set to 5500. All experiments were done on 4× NVIDIA RTX PRO 6000 Blackwell GPUs.

For data generation, 3 rollout branches were generated initially per case, 1 branch launch points are then chosen to branch. DTC needs no threshold as it is a relative metric. The RAC threshold is set to 3 as we found that the average RAC for correct samples in our teacher model Baichuan-M3 was 3.3. We also tried setting it to 4 which increased the data retention from 53% to 78% but yielded -3.5% final performance on Med-R.