Closing the Abductive Gap: Cognitive Feature Matching to Evaluate Hypotheses in Medical Diagnosis

Varul Srivastava [*1] **Narayan Hegde** [*1] **Mizhaan Maniyar** [1] **Aravindan Raghuveer** [1] **Vaibhav Rajan** [1]

Abstract

Diagnosis in realistic, open-ended clinical settings remains a challenge for Large Language Models (LLMs) as they struggle with abductive reasoning – the process of synthesizing patient observations to formulate the most plausible explanatory hypothesis. In this paper, we show that current approaches such as Chain-of-Thought prompting and general-domain reasoning metrics fail to improve diagnostic accuracy, and identify limitations in their ability to generate, evaluate, and rank hypotheses. To address this gap, we introduce the Abductive Reasoning Quality Score (ARQS), grounded in the psychologically principled Tversky's feature contrast model, to explicitly capture the cognitive structure of clinical judgement. We design strategies to employ ARQS across single- and multi-turn environments. Experiments on benchmark datasets demonstrate that ARQS-based instructions outperform existing strategies for clinical abduction and process-reward models improving both accuracy and robustness of LLMs. ARQS also serves as a discriminative reward signal for Reinforcement Learning, improving accuracy by up to 26%. Finally, in multi-turn simulated clinical environments, our ARQS-based protocol achieves superior diagnostic accuracy while reduced cost-of-care by reducing unnecessary medical testing.

1. Introduction

Large Language Models (LLMs) have achieved remarkable performance, frequently matching or surpassing human expert performance across a variety of medical benchmarks (Jin et al., 2019, 2021; Pal et al., 2022; Singhal et al., 2023a). However, recent in-depth evaluations reveal underlying fragility and critical reasoning gaps. For instance, adversarial stress tests have exposed their reliance on shortcut learning and rote memorization rather than genuine comprehension, failing when clinical prompts are trivially altered (Gu et al., 2025) and leading to a loss of generalization for novel tasks (Wu et al., 2023). Furthermore, these models lack essential metacognition, which includes awareness of one's own cognitive processes, e.g., bias identification, self-evaluation and monitoring (Griot et al., 2025). As a result, their performance is limited in realistic clinical settings that require flexible, open-ended reasoning and information-seeking strategies (Williams et al., 2024; Li et al., 2024; Pfohl et al., 2024), and fails to transfer safely to complex, real-world clinical problem-solving (Tu et al., 2025; Moëll et al., 2025; Sim & Chen, 2025).

To understand the underlying causes of these failures, we examine clinical diagnostic reasoning through the lens of Peirce's cycle of scientific inquiry, which integrates three distinct logical paradigms: abduction, deduction, and induction (Stanley, 2019; ten Cate & Durning, 2018). In both medical AI and broader Natural Language Processing (NLP) research, deductive reasoning (applying established medical knowledge to predict outcomes) and inductive reasoning (validating diagnostic hypotheses and generalizing patterns through repeated empirical observation) have been studied (Mondorf & Plank, 2024; Zhang et al., 2025; Sim & Chen, 2025). However, *abductive reasoning* – the indispensable first step of synthesizing noisy, incomplete patient observations to formulate the single most plausible explanatory hypothesis (Pietarinen & Stanley, 2025) – remains under-explored in clinical AI literature. Further, general approaches for abductive reasoning in non-medical domains, e.g., in abstract logic puzzles, code generation, and narrative inference (Del & Fishel, 2023; Sun et al., 2025; Bhagavatula et al., 2020; He et al., 2025), often lead to unconstrained hypothesis diversity which are not always clinically relevant (Gouveia, 2024; Moëll et al., 2025). The role of abduction in clinical decision-making is well established: it aligns with the clinical process of differential diagnosis enhancing interpretability, its process of iterative multiple hypothesis generation makes it robust to noise and incompleteness; and guides realistic multi-turn scenarios (Sim & Chen, 2025; Gouveia, 2024; Martini, 2023).

[1]Google Deepmind. Correspondence to: Varul Srivastavaa <varul@google.com>, Narayan Hegde <hegde@google.com>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

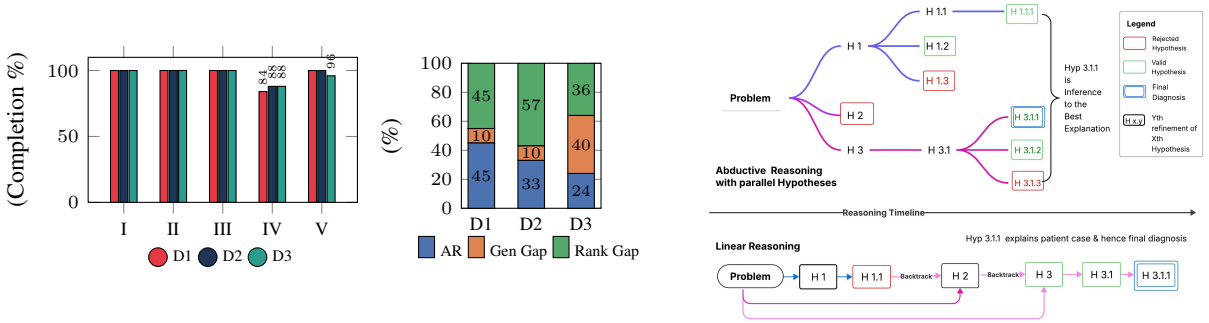


Figure 1. [**Left**] Performance of Gemini-3 Flash in each of the 5 stages of abductive reasoning across 3 datasets: only stage IV (hypothesis refinement) underperforms; [**Middle**] Diagnosis Accuracy of Gemini-3 Flash when instructed to think with Abductive Reasoning (AR); increase in accuracy when the Ground Truth is injected into the hypothesis set that the LLM generates before ranking them to produce final diagnosis (Gen Gap); improvement headroom due to inadequate hypothesis ranking (Rank Gap); [**Right**] Clinical abduction requires stage-wise generation of multiple plausible hypotheses with iterative refinement and their evaluation (above) while current LLMs reason about a single hypothesis in each iteration. D1: MedCaseReasoning (Wu et al., 2025), D2: DiReCT (Wang et al., 2024a), D3: MedXpertQA (Zuo et al., 2025). Experiment details are in sec. 4.

In this paper, we systematically investigate the reasoning capabilities of frontier LLMs across the distinct stages of the abductive clinical pipeline, which sequentially involves: (1) *observation grounding*, where salient symptoms and signs are extracted from a clinical presentation; (2) *hypothesis generation*, the initial proposition of a differential diagnosis; (3) *evidence-hypothesis mapping*, wherein the patient's specific data is aligned with the expected disease profiles; (4) *hypothesis refinement*, the iterative process of ruling conditions in or out based on missing data or clinical contradictions; and ultimately (5) *hypothesis ranking*, where the refined list is prioritized based on biological likelihood and immediate patient risk (Bowen, 2006; Patel et al., 1994). Our experiments on 3 benchmark datasets with a frontier model, instructed to think abductively via Chain-of-Thought (CoT) prompting (Wei et al., 2022), isolate the root cause of the problem in hypothesis generation and ranking (see fig. 1 and sec. 4). The underlying limitation essentially stems from inadequate abductive reasoning, esp. the ability to generate and evaluate multiple plausible *simultaneously* hypotheses iteratively, as opposed to the more common linear reasoning chains, common in non-medical domains (Guo et al., 2025), which generate and evaluate hypothesis one at a time.

To address this limitation, we develop a novel approach for scoring abductive reasoning and hypothesis quality using Amos Tversky's Feature Contrast Model (Tversky, 1977). Standard methods of scoring hypothesis quality often rely on geometric embeddings (e.g., cosine similarity) to match patient observations with disease representations. However, metric spaces rigidly enforce symmetry, minimality, and the triangle inequality – axioms that inherently contradict the asymmetric cognitive structure of clinical judgment (Appendix B). Tversky's model, instead, represents objects as discrete sets of features and measures similarity as an asymmetric linear combination of their shared (explanatory) and distinctive (missing or contrasting) features. Tversky's similarity has been shown to predict human categorization and judgment significantly more accurately than spatial metric measures in various empirical studies on human psychology (Goldstone, 1999). We posit that this also faithfully models the explicit feature-matching process utilized by expert physicians, as discussed in (Bordage, 1999); and the matching, missing and contrasting features across the sets of patient observations and disease profiles operationalize the heuristics of clinical abductive reasoning (Peng & Reggia, 1990) in an optimal way for differential diagnosis (see Appendix C).

Building upon this cognitive framework, we introduce the Abductive Reasoning Quality (ARQ) scoring of clinical hypotheses and design strategies for its use in single- and multi-turn settings. We illustrate its utility on benchmark datasets – as an instruction strategy during inference it addresses the limitation of inadequate ranking of hypotheses (*Rank Gap* in fig. 1); further, when used as a reward signal during reinforcement learning (RL), it improves overall performance including hypothesis generation (*Gen Gap* in fig. 1). This utility extends to multi turn diagnosis setting with reduces cost-of-care by eliminating unnecessary clinical investigations. Further, in adversarial tests designed to measure robustness to critical clinical details for diagnosis, we also show that ARQ enables reliance on actual evidence rather than spurious correlations, thereby enabling more trustworthy clinical AI. Through the use of an orchestrator (Nori et al., 2025), which uses clinical cases to simulate a real world clinical environment, we show that in multi-turn settings our approach leads to improved diagnostic accuracy with reduced cost-of-care by eliminating unnecessary

clinical investigations.

In summary, our contributions in this paper are:

- We conduct a systematic analysis of frontier LLMs on diagnosis tasks which reveals key limitations of hypothesis generation, evaluation and ranking in their abductive reasoning capabilities.
- We introduce and evaluate an explicit instruction strategy for clinical abductive reasoning, where LLMs generate and comparatively rank multiple hypotheses to arrive at a final diagnosis.
- We develop a novel metric, called ARQS, to measure the abductive reasoning quality of a diagnostic hypothesis, where clinical reasoning is modeled through a psychologically principled feature matching method. We design strategies to employ ARQS in single- and multi-turn abductive reasoning instruction prompting protocols and as a reward signal for RL training.
- Our extensive experiments, on single-turn and multi-turn diagnoses, demonstrate that ARQS outperforms state-of-the-art approaches for clinical abductive reasoning, in final diagnostic accuracy and robustness, both in inference-only as well as supervised RL settings.

2. Related Work

LLMs have achieved expert-level performance on medical benchmarks (e.g., MedQA(Jin et al., 2021), MedMCQA(Pal et al., 2022), PubMedQA(Jin et al., 2019)) and multimodal datasets, frequently matching human clinicians in constrained scenarios (Zhou et al., 2025; Thirunavukarasu et al., 2023; Nazi & Peng, 2024; Singhal et al., 2023a,b; Liu et al., 2025). This success is largely driven by general reasoning advancements evaluated in mathematics, coding, and generalization tasks (Cobbe et al., 2021; Hendrycks et al., 2021; Chen et al., 2021; Jimenez et al., 2024; Chollet, 2019; Mondorf & Plank, 2024). Techniques simulating deliberate "System 2" thinking (Kahneman, 2011)—such as Chain-of-Thought (CoT) (Wei et al., 2022), Self-Consistency (Wang et al., 2023), Tree and Graph of Thoughts (Yao et al., 2024; Besta et al., 2023), and LLM-as-a-Judge evaluations (Zheng et al., 2024)—have been widely adopted. In clinical settings, these deductive methods are increasingly enhanced by step-by-step verification using Process Reward Models (e.g., Med-PRM(Yun et al., 2025)) and others (Lightman et al., 2023; Golovneva et al., 2022).

Recent evaluations reveal that LLM outputs merely exhibit an "abductive appearance" driven by stochastic pattern-matching rather than genuine comprehension (Floridi et al., 2025; Sim & Chen, 2025). Clinical stress tests expose profound vulnerabilities: models rely heavily on shortcut learning and fabricate reasoning when inputs are perturbed, e.g., (Gu et al., 2025) show that frontier LLMs successfully guessed diagnoses even when critical evidence (such as medical images) was withheld, and failed catastrophically when text was adversarially altered to contradict the image, often fabricating reasoning to justify their incorrect priors. Moreover, they lack essential metacognition and uncertainty management (Griot et al., 2025), and suffer from inflexible, rote learning that fails on atypical cases (Kim et al., 2025). Furthermore, enhancing general reasoning capabilities does not seamlessly transfer to clinical problem-solving; fine-tuning often degrades actual reasoning quality, revealing that step-by-step logic in medical LLMs is frequently an opaque retrieval of memorized text rather than transparent clinical deduction (Tu et al., 2025; Moëll et al., 2025).

While standard benchmarks often treat diagnosis as single-turn question-answering, real-world diagnosis is a longitudinal, multi-turn process requiring the dynamic exploration of multiple hypotheses as new evidence emerges, as shown in multi-turn simulated environments like AMIE (Tu et al., 2024) and AgentClinic (Schmidgall et al., 2024) to structure diagnostic dialogue. Recent studies on multi-turn settings reveal severe limitations in LLMs, including fixation on initial hypotheses, inappropriate test ordering, premature closure, and failures to screen for critical conditions (Nori et al., 2025; Chiu et al., 2025); but they have not leveraged structured abductive reasoning to mitigate these errors.

LLMs struggle with abduction even in simple non-medical environments (Sun & Saparov, 2025). While general-domain frameworks like GEAR (He et al., 2025) evaluate abduction via logical consistency and unconstrained hypothesis diversity, they fail to translate safely to medicine (see Appendix D.3, where unconstrained generation risks biologically absurd diagnoses and violates strict patient safety constraints (Gouveia, 2024; Moëll et al., 2025). Although recent clinical literature has proposed aligning models with diagnostic thought processes, simulating forward-inference, and evaluating diagnostic uncertainty (Xu et al., 2024; Wang et al., 2024b; MediQ Project, 2024), these works neither fully evaluate nor intrinsically improve clinical abductive reasoning capabilities, leaving abductive hypothesis generation as a critical frontier for clinical AI.

3. The ARQS Framework for Scoring Abductive Reasoning Quality

We now present our scoring technique which plays the key role in our strategies for abductive reasoning in both single- and multi-turn settings. Let O denote the set of clinical observations found in a patient record. Let H denote a hypothesis or disease profile defined by the prototypical set of expected observations that characterizes it.

Tversky's Similarity Matching. A standard approach to such a scoring is to measure the degree of match be-

tween these two sets, O and H, e.g., by calculating the distance between geometric embeddings of the observations and the hypothesis (e.g., using Euclidean or cosine similarity). However, matching based on geometric embeddings fails to capture the cognitive structure of human clinical judgment, because geometric spaces rigidly enforce the three axiomatic properties of metric distances—symmetry, minimality, and the triangle inequality—all of which are violated in clinical reasoning (Tversky, 1977); see Appendix B. To accurately reflect clinical reasoning, we instead adopt Tversky's feature-matching process (Tversky, 1977), where objects are represented explicitly as discrete sets of features, and their similarity is measured as a linear combination of the measures of their shared and unique/contradictory features. This aligns with how expert physicians actually diagnose: by actively weighing semantic matches, noting the absence of expected findings, and scanning for clinical "red flags" rather than computing spatial distances (Bordage, 1999). We can formally deconstruct the relationship between the clinical observations (O) and the disease profile (H) into four distinct logical sets:

- **Agreement and Explanatory Coverage ($S_{Agree} = O \cap H$):** Observations which are present in the patient as well as in the disease profile represent positive clinical matches. In abductive reasoning, this set rewards a hypothesis for its explanatory power; the more patient findings the disease profile successfully covers, the stronger the abductive inference (Peng & Reggia, 1990; Josephson & Josephson, 1996).

- **Contradiction and Falsification ($S_{Contra} = O \cap \neg H$):** Set $\neg H$ represents the "forbidden" states of the hypothesis and this *Constrast Set* ($O \cap \neg H$) formalizes the abductive necessity of falsification. In clinical reasoning, a strong contradiction can outright invalidate a hypothesis (Bowen, 2006).

- **Missing Features and Iterative Testing ($S_{Miss} = H \setminus O$):** These features are expected by the disease profile and are absent from the patient's current observations. Such missing features actively dictate the next steps in patient care (e.g., ordering targeted labs to seek the missing data) (Bowen, 2006). It reflects the abductive hesitation to prematurely commit to a diagnosis without sufficient evidence.

- **Unexplained Findings and Parsimony ($S_{Unexp} = O \setminus H$):** These features are found in the patient that the proposed hypothesis cannot account for and could indicate the gap in diagnostic parsimony (Occam's Razor) (Patel et al., 1994). If severe symptoms remain unexplained, it prompts the search for a broader unifying diagnosis or a concurrent disease process.

Tversky's equation for similarity S balances features that are *shared* and *different*:

$$S(O, H) = \frac{f(O \cap H)}{[f(O \cap H) + \alpha f(O \cap \neg H) + \beta f(H \setminus O) + \gamma f(O \setminus H)]} \quad (1)$$

Here, $f(\cdot)$ represents a salience or weighting function for the features within a given set, and α, β, and γ are tunable parameters that dictate the penalty applied to different types of mismatches. By allowing for severe penalization of clinical "red flags" (α) and asymmetrical weighting of missing versus unexplained data ($\beta \neq \gamma$). See Appendix C for theoretical results on the optimality of S, demonstrating its suitability for differential diagnosis.

3.1. ARQS: Abductive Reasoning Quality Scoring

We operationalize clinical abduction into a structured, four-stage ARQS prompting system as shown in Algorithm 1 (more details are in Appendix E and prompts in Appendix H).

- **Stage 1 [Observation Grounding & Hypothesis Generation]** extracts key observations (O) from the input record and generates hypotheses to form a diverse, comprehensive differential diagnosis

- **Stage 2 [Clinical Profile Generation & Salience Calculation]** constructs an *ideal* clinical profile (H_i) for each candidate *disease* and explicitly compute feature *salience* ($f_i(\cdot)$). This salience is computed by prompting the model to assign an importance weight to every expected symptom, lab result, and precondition based on established medical knowledge (from 1.0 for pathognomonic features, to 0.3 for non-specific findings, and 0.0 for irrelevant data).

- **Stage 3 [ARQ Scoring]** computes the Tversky sets and the similarity scores based on Eq. 1 using O from stage 1 and H_i, f_i from stage 2. We also separately add an **atypical** set which is unusual for the diagnosis but not strictly contradictory. We apply an aggressive penalty coefficients to fatal clinical contradictions ($\alpha = 3.0$) and moderate penalties to missing or unexplained clinical noise ($\beta = 0.5, \gamma = 0.5$) for robust diagnosis. The score for each hypothesis is adjusted by a gentle epidemiological prior to appropriately anchor the probabilities of common conditions without overwhelming the patient-specific evidence.

- **Stage 4 [Hypothesis Selection]** selects the inference to the best explanation by performing an $\arg\max$ operation over the scores in stage 3, which the LLM is instructed to use for diagnosis. To ensure clinical safety, a subsequent postprocessing step incorporates

Algorithm 1 ARQS: Abductive Reasoning Quality Scoring

Input: Clinical scenario O_{raw}
Output: Diagnosis H_{final}
$O, \mathcal{H} \leftarrow$ GroundObservations(O_{raw}), GenerateHypotheses(O) `// Stage 1`
foreach $H_i \in \mathcal{H}$ **do**
  $P_i, f_i, B_i \leftarrow$ BuildIdealProfile(H_i), AssignWeights(P_i), EstimatePrior(H_i) `// Stage 2`
  $S_{set} \leftarrow$ MapTverskySets(O, P_i) for $set \in \{Agree, Contra, Atyp, Miss, Unexp\}$ `// Stage 3`
  $WC_{set} \leftarrow \sum f_i(S_{set})$; $S_{final}(H_i) \leftarrow WC_{Agr} \cdot \exp(-\sum WC_{others \neq Agr}) \cdot B_i^{0.3}$
$H_{final} \leftarrow LLM(\arg\max_{H_i} S_{final}(H_i))$ `// Stage 4`
return H_{final}

an override mechanism for rare cases where definitive, irrefutable clinical evidence (e.g., pathognomonic biopsies or specific PCR results) ought to be applied without considering numerical ranking.

We discuss how this quality scoring can be extended to a list of hypotheses to measure the quality of a differential diagnoses in Appendix E.2.

3.2. ARQS for Multi-Turn Reasoning

We now extend the foundational ARQS framework for dynamic, multi-turn environments, based on MAI-DxO (Nori et al., 2025), a model-agnostic orchestrator that simulates multiple doctors, proposes potential differential diagnoses and chooses high-value, cost-effective tests. In our setting, we have an iterative process with two primary components: (i) The **Simulator**: Acting as the environment, this model provides real-world feedback by mimicking patient responses, delivering test results, and detailing procedural outcomes. It evaluates the agent's performance through a reward system based on diagnostic accuracy and the cost-effectiveness of requested tests. The simulator's responses are grounded in patient case reports from the dataset used (MedCaseReasoning and MedXpertQA). (ii) The **Diagnosis Agent**: This model navigates the diagnostic journey turn-by-turn, utilizing a reasoning framework to systematically guide the interaction toward a conclusive diagnosis. The Diagnosis Agent strives to identify the correct condition within the fewest possible turns. Once it establishes a final hypothesis, it initiates a termination call to confirm the accuracy of its prediction. The reasoning framework can be simple CoT or based on abductive reasoning (see prompts in Appendix H).

We now describe the Multi-Turn ARQ Protocol (Algorithm 4), for such diagnostic agents. By continuously maintaining and updating the four Tversky sets across each conversational turn, it creates a turn-by-turn abductive loop comprising five stages: (1) **Integrate**, where new clinical data or test results are merged with the cumulative patient history; (2) **Falsify (Drop)**, wherein the model evaluates the Contradiction Set (Set B) to permanently discard hypotheses

Algorithm 2 Multi-Turn Dynamic ARQ System for Clinical Diagnosis

Input: Initial patient presentation $O_{initial}$
Output: Definitive final diagnosis H_{final}
$O_{cumul} \leftarrow O_{initial}$
$\mathcal{H}_{active} \leftarrow$ GenerateBroadDifferential(O_{cumul}) `// Include common, must-not-miss, and mimics`
while *Conversation is not Terminated* **do**
  `// 1. Integrate new data`
  O_{cumul}, PendingTests $\leftarrow$ IntegrateNewEvidence$(O_{cumul}$, ReceiveInput()$)$
  `// 2. ARQ Mapping & Falsify (Drop via Set B)`
  $\mathcal{H}_{active}, \mathcal{H}_{dropped} \leftarrow$ DropContradictedHypotheses$(\mathcal{H}_{active}, O_{cumul})$ `// +1.0 Reward`
  `// 3. Generate (Expand via Set D)`
  $\mathcal{H}_{active} \leftarrow \mathcal{H}_{active} \cup$ GenerateHypothesesForUnexplainedNoise$(O_{cumul}, \mathcal{H}_{active})$ `// -2.0 Penalty`
  `// 4. Retain & Rank (via Set A)`
  $H_{top} \leftarrow$ RankByBestExplanatoryFit$(\mathcal{H}_{active}, O_{cumul})[0]$
  `// 5. Terminate Check (Inference to Best Explanation)`
  if *IsDefinitiveMatch*(H_{top}, O_{cumul}) ***and*** *PendingTests* $= \emptyset$ **then**
    return Terminate(H_{top}) `// +20.0 Reward`
  `// 6. Investigate (Ask/Test via Set C) & Output`
  $Action \leftarrow$ SelectCostAwareAction(HighestYieldMissingFeatures(H_{top})) `// Incur tool costs`
  ExecuteActionAndFormatResponse$(Action, \mathcal{H}_{active}, H_{top})$

that face hard clinical contradictions; (3) **Generate (Expand)**, where the Unexplained Set (Set D) is analyzed to propose new hypotheses if significant patient symptoms remain unaddressed; (4) **Retain & Rank**, which re-evaluates the Agreement Set (Set A) to establish the current leading differential; and (5) **Investigate**, where the model evaluates the Missing Set (Set C) to identify the highest-yield information gap, driving the selection of the next patient question or diagnostic tool call to rule conditions in or out.

To mirror real-world competing constraints of diagnostic certainty and financial and physical costs of excessive medical testing, the agent's performance across multiple turns is evaluated using a cost-aware reward system (summarized in Table 4 in Appendix E), which incentivizes final diagnostic accuracy and efficient hypothesis pruning, while penalizing the overuse of invasive/expensive diagnostic tools and the unconstrained redundant hypothesis generation. Algorithm 2 summarizes the algorithm (more details are in Appendix E and prompts in Appendix H.

4. Experimental Results

4.1. Settings

Datasets. We use 3 benchmark datasets for clinical diagnosis, which provide clinical reasoning leading to the final diagnosis: MedCaseReasoning (Wu et al., 2025), MedXpertQA (Zuo et al., 2025) and DiReCT (Wang et al., 2024a). MedCaseReasoning, created from case reports, contains 14,489 diagnostic question-and-answer cases, with a test set of 897 cases, each paired with detailed reasoning statements,

including the differential diagnoses. DiReCT contains 511 EMR excerpts from the MIMIC-IV database (Johnson et al., 2023). across 25 disease categories. MedXpertQA contains expert-level questions across multiple medical specialties from diverse sources including medical license examinations. We use 1050 cases from the Text only subset. We converted the multiple-choice questions into open-ended diagnosis tasks and excluded the answer options, thereby preventing them from biasing the hypothesis generation process, and ensuring evaluation consistency across datasets.

Models. We evaluate clinical abductive reasoning capabilities across different models architectures and sizes. Among frontier closed-source models, we use Gemini-3.1 Pro, Gemini-3 flash and GPT-5 mini. Using both Gemini-3.1 pro and Gemini-3 flash allows us to see performance across model sizes, from the same family while Gemini-3 flash and GPT 5 mini allows comparing across model architectures among frontier models. Among open-weight models, we compare performance between Gemma-4 4B and Qwen-3 4B models. We also show RL experiments on these open source models. We disable thinking across open-weight models to (1) limit the rollout lengths for training feasability and (2) ensure consistency of process across models. While standard benchmarking uses greedy decoding, we apply a 0.5 inference temperature across models to promote diverse differential diagnoses. For GRPO (Shao et al., 2024)-based RL, we use a sampling temperature of 1.0 to ensure exploration and rollout variance. Additionally, we disable native latent thinking in open-weights models to maintain training feasibility and enforce adherence to the reasoning frameworks we evaluate.

4.2. Results

Diagnosing the Diagnostician: Why LLMs Fail at Clinical Abductive Reasoning. We investigate the abductive reasoning (AR) capability of a Gemini3-Flash through two experiments. We ask the LLM to think abductively (see prompts in Appendix H.1), by generating and ranking hypotheses before predicting the final diagnosis. First, we evaluate whether AR prompted LLM follows the 5 required AR steps (described in sec. 1) without judging the clinical correctness of each step: (i) Observation Grounding (ii) Hypothesis Generation (iii) Evidence-Hypothesis Mapping (iv) Dynamic Hypothesis Refinement (v) Comparative Ranking, where each is framed as a yes/no question. Another LLM-as-a-judge is used to evaluate the LLM traces. Fig. 1 (left) shows the performance in each step of AR on all 3 datasets. We observe that the model obtains perfect 100% score on all tasks except the task of Dynamic Hypothesis Refinement indicating the ease of following AR in diagnosis.

Second, we inject the Ground Truth (GT) diagnosis into the hypotheses set generated by the LLM before it ranks them to select the final diagnosis. Here, we measure the final diagnosis accuracy. Fig. 1 (middle) shows that this improves the final accuracy – to around 55% in MedCaseReasoning, 43% in DiReCT and 64% in MedXpertQA, but a significant headroom remains, motivating the problem of evaluating the hypotheses *vis-à-vis* the patient case to choose the final diagnosis.

Performance Comparison. We compare the performance of ARQS (Algorithm 1) with baselines: (i) **No AR**, the LLM is asked to directly give the final diagnosis, relying on it's internal CoT; (ii) **AR**: the LLM is asked follow abductive reasoning and provide the final diagnosis; (iii) **GEAR** (He et al., 2025) based score adapted for clinical scenarios (Appendix D); (iv) **MedPRM** (Yun et al., 2025), a process reward model designed for clinical domain; In (iii) and (iv), each generated hypothesis is scored and the highest scoring hypothesis is considered the final diagnosis. See Appendix H.2 for the prompts used.

Table 1 shows that ARQS outperforms the baselines in most cases. On MedCaseReasoning, the largest frontier model, Gemini 3.1 Pro works the best with No AR – this could be due to the constituent case reports being used in its pre-training, whereas AR-based prompts that provide structure to its thinking lead to reduced performance; among these ARQS has the best performance. To investigate the reason for its superior performance, we considered cases where the GT is part of the hypothesis set $h_{GT} \in H$, and computed the Normalized Marginal Gain $G = \frac{h_{GT} - h_{\text{top non GT}}}{h_{GT}}$ with respect to the *top-scoring non Ground Truth* hypothesis, $h_{\text{top non GT}}$. We find that ARQS has a higher gain compared to the next best baseline, indicating that it scores h_{GT} higher and $h_{\text{top non GT}}$ lower, as expected from a good scoring function. See Appendix G.1 for more details. Since ARQS generates continuous-valued scores (like Med-PRM), this high marginal gain underscores its strong discriminative power, making it an effective reward signal for RL based model optimization.

ARQ score as a reward signal. We run GRPO (Shao et al., 2024) to train models Qwen-3 4b Instruct (Yang et al., 2025) and Gemma-3 4b Instruct (Kamath et al., 2025). The model is given the patient case record P as input and instructed to output (1) list of hypotheses H and the final diagnosis $\hat{y}$. Let y denote the Ground Truth (GT) diagnosis. We compare the performance across two reward functions. (i) **Vanilla GT** ($R_{GT}(\hat{y}, y)$): computed by using a judge LLM (Gemini-3-flash) to match equivalence of predicted diagnosis and GT (see equivalence checking prompt in Appendix H.3); (ii) **medPRM** ($R_{\text{medPRM}}(H, P)$): We use the MedPRM distilled model to score reasoning trace for each hypothesis $h \in H$. The average score serves as the reward signal. (iii) **ARQS with GT**: We use the weighted sum $\alpha_{RL} \cdot R_{GT}(\hat{y}, y) + \beta_{RL} \cdot \sum_i R_{ARQS}(H_i, P)/|H|$ as the

Table 1. Performance Comparison of ARQS with baselines across models (row-wise best results in bold)

Dataset	Model	No AR	AR Prompt	GEAR	MedPRM	ARQS
MedCaseReasoning	Gemini-3 Flash	45.2 ± 1.1	43.4 ± 1.2	44.4 ± 1.01	39.8 ± 1.1	$\mathbf{45.6 \pm 0.8}$
	Gemini-3.1 Pro	$\mathbf{45.3 \pm 0.1}$	41.3 ± 1.0	36.0 ± 0.2	37.3 ± 1.0	44.5 ± 0.5
	GPT-5 mini	37.3 ± 1.2	39.7 ± 1.5	32.6 ± 0.7	38.2 ± 1.1	$\mathbf{42.2 \pm 0.7}$
	Qwen-3 4b	16.0 ± 0.5	20.2 ± 2.8	16.2 ± 1.3	15.3 ± 0.3	$\mathbf{23.2 \pm 0.3}$
	Gemma-4 4b	23.3 ± 0.6	$\mathbf{25.0 \pm 1.0}$	23.0 ± 1.8	20.5 ± 0.9	$\mathbf{25.0 \pm 0.7}$
DiReCT	Gemini-3 Flash	32.9 ± 1.1	27.98 ± 1.5	30.91 ± 0.5	36.98 ± 1.01	$\mathbf{40.11 \pm 0.8}$
	Gemini-3.1 Pro	38.6 ± 0.5	36.5 ± 0.2	27.6 ± 0.9	28.2 ± 0.3	$\mathbf{48.3 \pm 0.9}$
	GPT-5 mini	36.9 ± 2.0	34.1 ± 1.7	31.7 ± 0.8	31.8 ± 0.5	$\mathbf{37.5 \pm 0.3}$
	Qwen-3 4b	29.8 ± 1.2	34.0 ± 3.1	13.7 ± 1.9	23.8 ± 1.2	$\mathbf{34.8 \pm 0.3}$
	Gemma-4 4b	34.7 ± 0.8	31.8 ± 1.3	36.2 ± 1.2	33.3 ± 2.0	$\mathbf{36.5 \pm 0.5}$
MedXpertQA	Gemini-3 Flash	27.4 ± 1.5	35 ± 1.2	31.8 ± 0.8	28.8 ± 1.1	$\mathbf{36.2 \pm 0.5}$
	Gemini-3.1 Pro	35.9 ± 0.6	37.9 ± 1.2	35.9 ± 0.2	35.9 ± 0.4	$\mathbf{40.4 \pm 0.5}$
	GPT-5 mini	15.3 ± 1.4	15.3 ± 1.2	9.3 ± 2.0	21.2 ± 1.5	$\mathbf{28.3 \pm 1.4}$
	Qwen-3 4b	11.3 ± 3.0	9.2 ± 2.8	9.0 ± 5.9	15.0 ± 2.3	$\mathbf{15.3 \pm 2.0}$
	Gemma-4 4b	13.5 ± 1.0	$\mathbf{14.7 \pm 0.8}$	13.5 ± 1.0	11.0 ± 1.7	13.7 ± 1.4

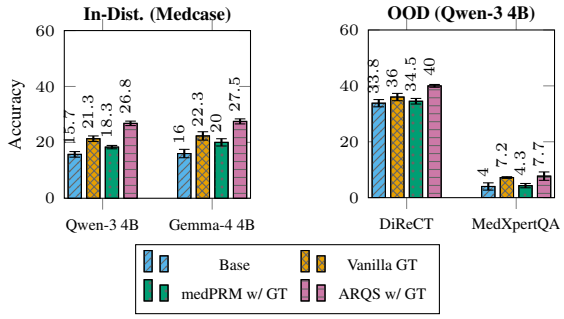


Figure 2. GRPO trained model performance across In-Distribution and Out-of-Distribution settings.

reward, where the average ARQ score is computed with Gemini-3-flash using the output H. Training is done using the train split of MedCaseReasoning. Results are averaged across 3 runs. We evaluate the diagnosis accuracy performance on the test split of MedCaseReasoning (In-Distribution) and on DiReCT and MedXpertQA (Out-of-Distribution). We compare the three Reward signals against the **Base** model (without any training).

Fig. 2 shows that ARQS with GT serves as a better reward function leading to 25.8% and 23.3% improvement in in-distribution accuracy in the two models, over using just GT as reward; In out-of-distribution settings, we observe 11.1% and 6.9% improvement over GT as reward for DiReCT and MedXpertQA respectively. MedPRM scores do not have the same effect and consistently reduce the performance compared to GT as reward. Accuracy progression over checkpoints and reward curves are in Appendix G. Overall, these results indicate that ARQS-based training leads to, at some level, the formation of higher quality hypotheses, which is a dataset-agnostic skill pertaining to System-2 thinking for clinical diagnosis.

Robustness to Adversarial Stress. We evaluate the robustness of LLMs in noisy environments in purely text-based reasoning, similar to multimodal adversarial tests in (Gu et al., 2025), across 3 instruction strategies No-AR, AR and ARQS with Gemini-3 Flash. To conduct this experiment, we isolate the "pivot" finding within a clinical vignette – the highly specific, anchoring clinical detail necessary to establish the Ground Truth (GT) diagnosis. These and the following multi-turn experiments do not utilize DiReCT as it does not contain clinical case reports and has relatively fewer patient observations and diagnoses. We generate adversarial variants for each clinical case as separate datasets (see Appendix H.4): (i) **Dataset A (Removed Pivot):** The pivot finding is systematically excised from the clinical vignette, rendering the case ambiguous. (ii) **Dataset B (Negated Pivot):** The pivot finding is explicitly negated (e.g., altering "positive for X" to "negative for X"), directly contradicting the GT diagnosis. Only valid cases upon negation are considered for evaluation. Clinical robustness requires that a model's diagnostic accuracy *decreases* when pivot findings are removed or negated. This shows that the model relies on the presented evidence rather than priors or spurious correlations, ensuring higher clinical safety. Additionally, models must also demonstrate *epistemic humility* by abstaining or declaring insufficient information when evidence is missing or contradicted.

Table 2 shows that ARQ is more robust to both negation and removal of pivots, compared to the baselines. We observe that pivot removal leads to anchoring to original GT diagnoses and pivot negation often leads to hallucinations of rare diseases to explain contradictions – both are addressed in ARQS. Negation leads to increase in contradictions, while removal leads to increase in missingness, which are reflected in the corresponding Tversky set-cardinalities and thereby in ARQS scores.

Variability. We evaluate diagnostic variability by measuring prediction consistency over five repeated samples per patient case. The results in Appendix G.3 show that compared to the No-AR and standard AR baselines, the ARQS framework demonstrates lower variability across all three datasets.

ARQ in multi-turn settings. We compare Multi-Turn Dynamic ARQ (Algorithm 2) with other reasoning strategies

Table 2. Robustness and Epistemic Humility

		Robustness (%) ↓			Abstain (%) ↑		
Benchmark	**Dataset**	**NO-AR**	**AR**	**ARQ**	**NO-AR**	**AR**	**ARQ**
MedCase	Negation	29.28	30.96	**26.61**	1.00	1.00	**2.00**
	Removal	25.55	23.43	**20.21**	26.00	22.50	**32.50**
MedXpertQA	Negation	17.60	21.60	**14.80**	5.00	2.00	**9.50**
	Removal	23.60	28.80	**21.60**	4.00	1.50	**8.00**

Table 3. Multi-Turn Diagnostic Performance across Prompting Strategies and Models

		Gemini 3.1 Flash-Lite			Gemini 3.1 Pro		
Dataset	**Metric**	**Baseline**	**AR**	**ARQ**	**Baseline**	**AR**	**ARQ**
MedCase	Accuracy ↑	32.00%	36.00%	**39.00%**	63.50%	64.00%	**65.30%**
	Final Reward ↑	0.89	0.93	**1.10**	1.541	1.72	**1.948**
	Avg. # Turns	7.38	6.81	7.69	4.88	5.91	8.66
MedXpertQA	Accuracy ↑	36.05%	38.10%	**40.75%**	47.92%	49.10%	**51.05%**
	Final Reward ↑	-0.4978	-0.3314	**0.0212**	1.3347	1.4667	**1.4960**
	Avg. # Turns	8.20	6.37	7.87	2.53	2.17	3.92

within a multi-turn environment described in sec. 3.2 In **baseline**, the model is instructed to iteratively alternate between asking patient questions and ordering diagnostic tests until it reaches a final diagnosis, with no guidance on how to structure its internal logic; in **AR**, the model is instructed to follow a 5-step abductive reasoning protocol in each turn: summarizing observations, listing differential diagnoses with confidence scores, dynamically evolving hypotheses, synthesizing case understanding, and planning actions to maximize information gain. All approaches are evaluated using a reward that penalizes excessive testing and hypothesis generation while rewarding diagnostic accuracy and efficient hypothesis pruning (see Appendix H.5 for the prompts).

Table 3 shows that ARQ outperforms the baselines in both final reward and diagnostic accuracy while maintaining a comparable value in average number of turns in most cases. We observe how Tversky's sets lead to this improvement in several cases (examples in Appendix G.4). The baseline often fixates on the first salient diagnosis which AR counters by requiring a differential diagnosis. ARQS prevents this anchoring bias and early fixation by tracking Set B (Contradictions) and dropping hypotheses falsified by new evidence. Premature closure is also avoided by requiring high agreement (Set A), and low unexplained observations (Set D). The baseline often jumps prematurely to advanced testing before gathering sufficient history. AR prevents this by requiring a step that justifies the need for an action and ARQS optimizes by explicitly linking next-step tests to Set C (Missing), ensuring actions target high-yield information gaps. Baseline models routinely omit rare but fatal etiologies, which AR necessitates in the initial differential. ARQS strengthens this safety net by continually monitoring Set D (Unexplained) and generating new hypotheses when required.

5. Conclusion

In this work, we demonstrated that state-of-the-art LLMs struggle to formulate, evaluate, and rank parallel diagnostic hypotheses, and that standard reasoning strategies (like CoT) are insufficient to bridge this gap. We designed a psychologically principled ARQ Score for diagnostic hypotheses, and single- and multi-turn instruction prompting protocols to effectively operationalize clinical abduction. Our empirical results demonstrate that ARQS outperforms extant abductive reasoning (AR) strategies across both single- and multi-turn clinical settings for diagnosis tasks, provides valuable discriminative reward signal for RL, and enhances epistemic humility and robustness in LLMs against noisy and incomplete data.

Limitations and Future Work: Our experiments have relied solely on the LLM's internal knowledge base (KB) for the crucial steps of generation of hypotheses and their clinical profiles. The effect of external KBs as well as agentic environments can be investigated further in the context of AR. Our solution to address the gap of hypothesis generation was via RL. We envisage alternative approaches independent of the scoring mechanism as an interesting open problem. Our implementation of Tversy scoring (Eq. 1) uses static dataset-specific hyperparameters. Developing dynamic, disease-specific weighting which are potentially more realistic would be another avenue of research. Finally, additional human evaluation of reasoning traces from ARQS and competing methods will be required to ascertain its clinical interpretability and actionability.

References

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., Lehmann, T., Podstawski, M., Niewiadomski, H., Nyczyk, P., et al. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*, 2023.

Bhagavatula, C., Le Bras, R., Malaviya, C., Sakaguchi, K., Holtzman, A., Hannah, G., and Choi, Y. Abductive commonsense reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

Bordage, G. Why did i miss the diagnosis? some cognitive explanations and educational implications. *Academic Medicine*, 74 (10):S138–S143, 1999.

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

Cabitza, F., Rasoini, R., and Gensini, G. F. Unintended consequences of machine learning in medicine. *JAMA*, 318(6): 517–518, 2017.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Chiu, C., Pitis, S., and van der Schaar, M. Simulating viva voce examinations to evaluate clinical reasoning in large language models. *arXiv preprint arXiv:2510.10278*, 2025.

Chollet, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Del, M. and Fishel, M. True detective: A deep abductive reasoning benchmark undoable for gpt-3 and challenging for gpt-4. *arXiv preprint*, 2023.

Floridi, L. et al. What kind of reasoning (if any) is an llm actually doing? on the stochastic nature and abductive appearance of large language models. *arXiv preprint arXiv:2512.10080*, 2025.

Ghassemi, M., Oakden-Rayner, H., and Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 4(11):e745–e750, 2021.

Goldstone, R. L. Similarity. In Wilson, R. A. and Keil, F. C. (eds.), *The MIT Encyclopedia of the Cognitive Sciences*, pp. 763–765. MIT Press, Cambridge, MA, 1999.

Golovneva, O. et al. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*, 2022.

Gouveia, S. S. Crossing the trust gap in medical ai: building an abductive bridge for xai. *Philosophy & Technology*, 37(3), 2024.

Griot, M. et al. Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16:642, 2025.

Gu, Y. et al. The illusion of readiness in health ai. *arXiv preprint arXiv:2509.18234*, 2025.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

He, K. et al. Gear: A general evaluation framework for abductive reasoning. *arXiv preprint arXiv:2509.24096*, 2025.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Jimenez, C. E., Murphy, J. Y., Shirikjian, V., Rosenberg, A., Mckay, W., Han, J., et al. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, 2024.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. 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.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Josephson, J. R. and Josephson, S. G. *Abductive inference: Computation, philosophy, technology*. Cambridge University Press, 1996.

Kahneman, D. *Thinking, fast and slow*. Macmillan, 2011.

Kamath, G. T. A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ram'e, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.-B., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R. I., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A. M., Goedeckemeyer, A., Saade, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., Gyorgy, A., Pinto, A. S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C. L., Choquette-Choo, C. A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D. S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Pluci'nska, H., Singh, H., Mehta, H., Lehri, H. T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J. M., Lai, J., Orbay, J., Fernandez, J., Newlan, J., Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M.,

Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P. K., Culliton, P., Schmid, P., Sessa, P. G., mei Xu, P., Stańczyk, P., Tafti, P. D., Shivanna, R., Wu, R., Pan, R., Rokni, R. A., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S. R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Lo, J., Moreira, E., Martins, L. G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V. S., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Lepikhin, D., Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., and Hussenot, L. Gemma 3 technical report. *ArXiv*, abs/2503.19786, 2025. URL https://api.semanticscholar.org/CorpusID:277313563.

Kim, J. et al. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *arXiv preprint arXiv:2502.04381*, 2025.

Li, S. S., Balachandran, V., Feng, S., Ilgen, J. S., Pierson, E., Koh, P. W., and Tsvetkov, Y. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Liu, X. et al. A generalist medical language model for disease diagnosis assistance. *Nature Medicine*, 31(3):932–942, 2025.

Martini, C. Abductive reasoning in clinical diagnostics. In *Handbook of abductive cognition*, pp. 467–479. Springer, 2023.

MediQ Project. Mediq: Evaluating diagnostic uncertainty and hypothesis generation. https://stellalisy.com/projects/mediQ/, 2024.

Moëll, B. et al. Medical reasoning in llms: an in-depth analysis of deepseek r1. *Frontiers in Artificial Intelligence*, 8:1616145, 2025.

Mondorf, P. and Plank, B. Beyond accuracy: Evaluating the reasoning behavior of large language models – a survey. *arXiv preprint arXiv:2404.01869*, 2024.

Nazi, Z. A. and Peng, W. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, pp. 57. MDPI, 2024.

Nori, H., Daswani, M., Kelly, C., Lundberg, S., Ribeiro, M. T., Wilson, M., Liu, X., Sounderajah, V., Carlson, J., Lungren, M. P., et al. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Pal, A., Umapathi, L. K., and Sankarasubbu, M. 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.

Patel, V. L., Arocha, J. F., and Kaufman, D. R. Diagnostic reasoning and medical expertise. *Psychology of Learning and Motivation*, 31:187–252, 1994.

Peng, Y. and Reggia, J. A. *Abductive inference models for diagnostic problem-solving*. Springer-Verlag, New York, NY, 1990.

Pfohl, S. R., Cole-Lewis, H., Sayres, R., Neal, D., Asiedu, M., Dieng, A., Tomasev, N., Rashid, Q. M., Azizi, S., Rostamzadeh, N., et al. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, pp. 1–11, 2024.

Pietarinen, A.-V. and Stanley, D. E. The logic of medical reasoning: toward an integrated inductive, deductive, and abductive approach to clinical practices. *Philosophy, Ethics, and Humanities in Medicine*, 2025.

Schmidgall, S., Ziaei, R., Harris, C., Kim, J. W., Reis, E., Jopling, J., and Moor, M. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sim, S. Z. Y. and Chen, T. Critique of impure reason: Unveiling the reasoning behaviour of medical large language models. *eLife*, 14:e106187, 2025.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Ajay, A., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023a.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023b.

Stanley, D. E. The logic of medical diagnosis: Generating and selecting hypotheses. *Topoi*, 38(3):437–446, 2019.

Sun, Y. and Saparov, A. Language models do not follow occam's razor: A benchmark for inductive and abductive reasoning. *arXiv preprint arXiv:2509.03345*, 2025.

Sun, Z. et al. Eliciting and improving the causal reasoning abilities of large language models with conditional statements. *Computational Linguistics*, 2025.

ten Cate, O. and Durning, S. J. Understanding clinical reasoning from multiple perspectives: A conceptual and theoretical overview. In *Principles and Practice of Case-based Clinical Reasoning Education*, pp. 35–46. Springer, 2018.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

Tu, H. et al. Knowledge or reasoning? a close look at how llms think across domains. *arXiv preprint arXiv:2506.02126*, 2025.

Tu, T., Palepu, A., Schaekermann, M., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Tomasev, N., et al. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*, 2024.

Tversky, A. Features of similarity. *Psychological Review*, 84(4): 327–352, 1977.

Wang, B., Chang, J., Qian, Y., Chen, G., Chen, J., Jiang, Z., Zhang, J., Nakashima, Y., and Nagahara, H. Direct: Diagnostic reasoning for clinical notes via large language models. *Advances in neural information processing systems*, 37:74999–75011, 2024a.

Wang, G. et al. Fmt: Simulating clinical reasoning via forward-inference framework in medical dialogue systems. *IEEE Transactions*, 2024b.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., et al. Do no harm: a roadmap for responsible machine learning for healthcare. *Nature Medicine*, 25(9):1337–1340, 2019.

Williams, C. Y., Miao, B. Y., Kornblith, A. E., and Butte, A. J. Evaluating the use of large language models to provide clinical recommendations in the emergency department. *Nature Communications*, 15(1):8236, 2024.

Wu, K., Wu, E., Thapa, R., Wei, K., Zhang, A., Suresh, A., Tao, J. J., Sun, M. W., Lozano, A., and Zou, J. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*, 2025.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*, 2023.

Xu, K. et al. Reasoning like a doctor: Improving medical dialogue systems via diagnostic reasoning process alignment. *arXiv preprint arXiv:2406.13934*, 2024.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Yun, J., Sohn, J., Park, J., Kim, H., Tang, X., Shao, Y., and Kang, J. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. *arXiv preprint arXiv:2506.11474*, 2025.

Zhang, D., Li, Z.-Z., Zhang, M.-L., Zhang, J., Liu, Z., Yao, Y., Xu, H., Zheng, J., Chen, X., Zhang, Y., et al. From system 1 to system 2: a survey of reasoning large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Zheng, L., Chiang, W.-L., Ying, Y., Hao, S., Zhang, Y., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Zhou, S. et al. Large language models for disease diagnosis: a scoping review. *npj Artificial Intelligence*, 1(1), 2025.

Zuo, Y., Qu, S., Li, Y., Chen, Z., Zhu, X., Hua, E., Zhang, K., Ding, N., and Zhou, B. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *Forty-Second International Conference on Machine Learning (ICML)*, 2025.

A. Broader Impact

The integration of LLMs into clinical workflows carries profound societal implications. By structuring medical AI through our proposed Abductive Reasoning Quality Score (ARQS), this work makes a step toward transparent, psychologically grounded reasoning. Formalizing clinical heuristics intrinsically mitigates diagnostic errors like premature closure, advancing the critical mandate for safe, responsible machine learning in healthcare (Wiens et al., 2019). Furthermore, unlike post-hoc explainability (XAI) methods that can provide false or misleading confidence (Ghassemi et al., 2021), ARQS potentially offers inherent interpretability by explicitly mapping reasoning to discrete clinical sets, fostering trustworthy human-AI collaboration.

However, the deployment of such systems, in general, introduces significant societal risks which have been discussed in the clinical AI literature. For instance, ARQS utilizes epidemiological priors which, if derived from historically biased training data, risk perpetuating systemic health disparities and discriminating against marginalized populations (Obermeyer et al., 2019). Additionally, because LLMs (both with and without ARQS) produce highly structured and confident reasoning traces, it elevates the risk of automation bias. Over-reliance on these automated outputs could lead to the deskilling of clinicians and potential degradation of independent diagnostic reasoning over time (Cabitza et al., 2017). Furthermore, maintaining multiple hypotheses across multi-turn interactions incurs non-trivial computational and environmental costs due to increased inference overhead.

To safely realize the benefits of abductive medical AI, the underlying models must be rigorously audited for demographic bias. Ultimately, systems utilizing ARQS must be deployed strictly as assistive tools within a "human-in-the-loop" framework, ensuring that AI serves to augment clinical expertise rather than replace the physician in the diagnostic process (Topol, 2019).

B. Why Geometric Models are poor for Clinical Judgment

Standard metric similarities (e.g., Euclidean or cosine distance) fail to model clinical judgment because they rigidly enforce geometric axioms that contradict human cognition (Tversky, 1977). Clinical diagnosis fundamentally violates these three metric properties:

1. **Symmetry ($d(O, H) = d(H, O)$):** Diagnosis is inherently asymmetric. A patient presenting with isolated fatigue (O) might be loosely compared to the prototype of Leukemia (H). However, the Leukemia prototype (H)—which entails severe anemia, bruising, and hepatosplenomegaly—is dissimilar to mere fatigue (Tversky, 1977).

2. **Minimality ($d(x, x) = d(y, y) = 0$):** Metric spaces dictate that the distance between any identical pair is exactly zero, making all perfect matches mathematically equal. Clinically, however, a patient perfectly matching a highly complex, specific syndrome (e.g., Systemic Lupus) is cognitively judged as a "stronger" and more salient diagnostic match than a patient matching a vague condition (e.g., a mild viral infection). In human reasoning, rich feature overlap increases the magnitude of perceived similarity, a nuance minimality erases.

3. **Triangle Inequality ($d(A, C) \leq d(A, B) + d(B, C)$):** This axiom routinely fails in medicine. A patient with isolated right-lower quadrant pain (A) is highly similar to Appendicitis (B). Appendicitis (B), in turn, shares systemic inflammatory features (fever, leukocytosis) with Bacterial Pneumonia (C). Geometrically, the triangle inequality implies A must be somewhat similar to C; clinically, however, localized abdominal pain shares no meaningful diagnostic overlap with a respiratory infection.

Because expert physicians evaluate diagnoses through asymmetric, context-dependent feature-matching rather than rigid spatial distances (Bordage, 1999), Tversky's contrast model provides a superior mathematical framework for modeling clinical abduction (Goldstone, 1999).

C. Mathematical Optimality of the ARQS Diagnostic Model

C.1. Formulation of the Diagnostic Feature Space

Let Ω be the universal set of all possible discrete clinical features (e.g., symptoms, signs, laboratory values). A diagnostic encounter evaluates a Patient Observation ($O \subset \Omega$) against a Disease Hypothesis ($H \subset \Omega$). Furthermore, diseases strictly forbid certain biological states, defining a set of mutually exclusive features $\neg H \subset \Omega$, where $H \cap \neg H = \emptyset$.

Definition 1 (The ARQS Partition)**.** *We map the clinical encounter (O, H) into a 4-tuple of disjoint feature sets $\mathbf{X} = (X_1, X_2, X_3, X_4)$ that exhaustively categorize the evidence:*

1. ***Agreement (X_1):*** *$O \cap H$ (Supporting evidence)*

2. ***Contradiction (X_2):*** *$O \cap \neg H$ (Falsifying evidence)*

3. ***Missing (X_3):*** *$H \setminus O$ (Expected but unobserved evidence)*

4. ***Unexplained (X_4):*** *$O \setminus (H \cup \neg H)$ (Clinical noise / incidental findings)*

Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \times \mathcal{X}_4$ be the space of all such possible 4-tuples. Let $\succeq$ be a binary relation on $\mathcal{X}$ representing *clinical preference*. The expression $\mathbf{X} \succeq \mathbf{Y}$ signifies that the clinical evidence $\mathbf{X}$ provides "at least as strong a diagnostic match" as $\mathbf{Y}$.

C.2. Axiomatization of Clinical Preference

To derive the optimal function that *does* preserve $\succeq$, we establish the cognitive axioms governing how clinical evidence is evaluated. Drawing on the foundational axioms of additive conjoint measurement (Krantz et al., 1971), we assume $\succeq$ satisfies the following conditions on $\mathcal{X}$:

Axiom 1 (Weak Ordering)**.** *The relation $\succeq$ is transitive and connected. For any encounters $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathcal{X}$:*

- *If $\mathbf{A} \succeq \mathbf{B}$ and $\mathbf{B} \succeq \mathbf{C}$, then $\mathbf{A} \succeq \mathbf{C}$.*
- *Either $\mathbf{A} \succeq \mathbf{B}$ or $\mathbf{B} \succeq \mathbf{A}$.*

Axiom 2 (Joint Independence / Additivity of Evidence)**.** *The clinical impact of any subset of the components $\{X_1, X_2, X_3, X_4\}$ is independent of the fixed values of the remaining components. Clinical Interpretation: If we hold the Missing (X_3) and Unexplained (X_4) sets constant across two different patients, the physician's preference depends strictly on their Agreement (X_1) and Contradiction (X_2) sets. A coincidental rash (X_4) does not mathematically alter the absolute diagnostic weight of a positive biopsy (X_1).*

Axiom 3 (Restricted Solvability & Archimedean Condition)**.** *The feature space is sufficiently rich (dense) such that similarities can be matched across different domains, and no single piece of non-falsifying evidence is infinitely more powerful than a combination of others. This ensures the ordinal ranking can be mapped to the real numbers $\mathbb{R}$.*

Axiom 4 (Clinical Monotonicity)**.** *Let $\mathbf{X} = (X_1, X_2, X_3, X_4)$ and $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$. If $X_1 \supseteq Y_1$ (more agreement), $X_2 \subseteq Y_2$ (less contradiction), $X_3 \subseteq Y_3$ (less missing), and $X_4 \subseteq Y_4$ (less unexplained), then $\mathbf{X} \succeq \mathbf{Y}$.*

C.3. Step-by-Step Proof of the ARQS Additive Conjoint Representation

We now prove that the only valid mathematical function that preserves these axioms is the linear additive contrast model (the ARQS linear basis).

Step 4.1: The Additive Conjoint Measurement Theorem

By the Theorem of n-component Additive Conjoint Measurement (Krantz et al., 1971, Theorem 6.2), if a relation $\succeq$ on a product space $\mathcal{X} = \prod_{i=1}^{n} \mathcal{X}_i$ (where $n \geq 3$) satisfies Weak Ordering (Axiom 1), Independence (Axiom 2), and Solvability (Axiom 3), then there **exist** real-valued functions $\phi_i : \mathcal{X}_i \to \mathbb{R}$ for each component such that for all $\mathbf{X}, \mathbf{Y} \in \mathcal{X}$:

$$\mathbf{X} \succeq \mathbf{Y} \iff \sum_{i=1}^{4} \phi_i(X_i) \geq \sum_{i=1}^{4} \phi_i(Y_i) \tag{2}$$

Logical Implication: This theorem guarantees that because clinical features are evaluated independently (Axiom 2), the *only* valid way to compute total similarity is by mapping each set to a real number independently, and **summing** them. No complex, non-linear geometric projection can preserve this fundamental relation.

Step 4.2: Applying Monotonicity to the Functions ϕ_i

Let $f(X)$ be an extensive, non-negative measure of the feature set X (e.g., the sum of the saliency weights of the features in X, such

that $f(\emptyset) = 0$).

By Axiom 4 (Clinical Monotonicity):

- $\phi_1(X_1)$ must strictly increase as X_1 grows. Therefore, $\phi_1(X_1) = \theta f(X_1)$ for some constant $\theta > 0$.
- ϕ_2, ϕ_3, ϕ_4 must strictly decrease as X_2, X_3, X_4 grow. Therefore, they must take negative values:

$$\begin{aligned}\phi_2(X_2) &= -\alpha f(X_2)\\ \phi_3(X_3) &= -\beta f(X_3)\\ \phi_4(X_4) &= -\gamma f(X_4)\end{aligned}$$

  where $\alpha, \beta, \gamma \geq 0$.

STEP 4.3: CONSTRUCTING THE LINEAR ARQS MODEL

Substituting these into the additive structure from Step 4.1 yields the optimal linear mapping $V_{linear}(O, H)$:

$$V_{linear}(O, H) = \begin{array}{l}\theta f(O \cap H) - \alpha f(O \cap \neg H)\\ \quad - \beta f(H \setminus O) - \gamma f(O \setminus (H \cup \neg H))\end{array} \tag{3}$$

This mathematically mirrors Tversky's Contrast Model (Tversky 1977, p. 332, eq. ii), formally expanded for the four-set clinical reality.

C.4. Normalization to the ARQS Ratio Form

While V_{linear} perfectly preserves the ordinal ranking of hypotheses, modern computational systems and probabilistic diagnostic models require scores bounded between $[0, 1]$.

According to Tversky's matching function formulations (1977, p. 333), an additive contrast model can be equivalently expressed as a Ratio Model by normalizing the positive evidence by the total absolute evidence evaluated. Setting the support scalar $\theta = 1$, we divide the *Agreement* component by the linear sum of all active components:

$$Validity(O, H) = \frac{f(S_{Agree})}{\begin{array}{l}f(S_{Agree}) + \alpha f(S_{Contra})\\ \quad + \beta f(S_{Miss}) + \gamma f(S_{Unexp})\end{array}} \tag{4}$$

Because this transformation is a monotonic scaling of the additive conjoint structure, it perfectly preserves the ordinal relation $\mathbf{X} \succeq \mathbf{Y}$ derived in Section 4.

C.5. Conclusion of Optimality

We have demonstrated via rigorous mathematical proof that:

1. **Metric Spaces Fail:** The topological constraints of continuous vector embeddings (Symmetry, Triangle Inequality) mathematically contradict the asymmetry of human clinical evaluation.
2. **Additivity is Required:** By the axioms of conjoint measurement theory, any evaluation strategy that assesses independent pieces of evidence monotonically *must* be isomorphic to an additive linear structure.
3. **ARQS is the Unique Solution:** The ARQS ratio formula is not a heuristic; it is the unique mathematical mapping that preserves the ordinal preferences of expert clinical cognition, allowing for asymmetric penalization ($\alpha \neq \beta \neq \gamma$) of missing versus contradictory evidence.

Therefore, the ARQS model is mathematically proven to be the most optimal and theoretically sound method for measuring patient-disease similarity. ■

D. The GEAR Evaluation Framework

Traditional benchmarks evaluate abductive reasoning by scoring a single generated hypothesis against a static "gold" answer. To address the inherent underdetermination of abductive reasoning—where multiple valid hypotheses can explain the same evidence—the General Evaluation for Abductive Reasoning (GEAR) framework (He et al., 2025) introduces a label-free paradigm. GEAR evaluates an entire set of generated hypotheses, denoted as $F = \{f_1, f_2, \ldots, f_m\}$, based on a set of initial observations $\mathbb{O}$ and a broader sample space of unseen inputs S. Each observation $o_i \in \mathbb{O}$ is an input-output pair. A hypothesis is a function mapping inputs to outputs. E.g., $o_1 =$ (floor, wet) and the hypothesis 'rain' satisfies this observation $f_{\text{rain}}(\text{floor}) = \text{wet}$. The framework operationalizes abductive reasoning quality across three core dimensions: Consistency, Generalizability, and Diversity. We first provide their definitions following (He et al., 2025) and then discuss their utility for clinical reasoning.

1. **Consistency:** A fundamental prerequisite for any abductive hypothesis is that it must logically entail the immediate premises. Consistency measures the proportion of the generated hypothesis set F that correctly explains all the given observations in $\mathbb{O}$. Formally, a generated hypothesis f is consistent with the observation set $\mathbb{O}$ if $\forall(in_i, out_i) \in \mathbb{O}, f(in_i) = out_i$.
2. **Generalizability:** GEAR posits that a stronger hypothesis is one that yields meaningful, determinate predictions over a broader set of unseen inputs. Given a problem-specific sample space of unseen inputs S, generalizability is measured by the fractional coverage of S where the hypothesis f can make a valid prediction. Formally, if a hypothesis f_i operates over an input domain $I_i \subseteq S$, and μ represents a set-size measure (such as cardinality), the generalizability of f_i is proportional to the size of its valid input domain: $\text{Generalizability}(f_i) \propto \frac{\mu(I_i)}{\mu(S)}$. A hypothesis f_1 is strictly more generalizable than f_2 if $\mu(I_1) > \mu(I_2)$, meaning it avoids overfitting to the initial observations $\mathbb{O}$ and successfully extrapolates to a wider array of novel scenarios.
3. **Diversity:** To prevent an AI model from generating redundant or superficial variants of the same explanation, GEAR evaluates the semantic diversity of the hypothesis set $\mathbb{F} = \{f_1, f_2, \ldots, f_m\}$. Adapting concepts from ecological diversity theory, GEAR quantifies diversity based on the distinct prediction patterns the hypotheses produce over the sample space S. This is measured using two complementary metrics:
   - γ-diversity (Average Unique Predictions per Input): This metric evaluates how many genuinely distinct predictions the hypothesis set offers for a given input. Let

$P(f) = \{(\text{in}, f(\text{in})) : f \in \mathbb{F}\}$, the set of all predictions made by hypothesis f for an input $\text{in} \in S$. The γ-diversity is the average cardinality of P_f:

$$\gamma\text{-diversity}(\mathbb{F}; S) = \frac{1}{|S|} \cup_{f \in \mathbb{F}} |P_f|$$

A higher γ score indicates that the hypotheses in F lead to fundamentally different outcomes and interpretations of the data.

- β-diversity (Prediction-Pattern Dispersion): This metric assesses the pairwise dissimilarity of the hypotheses across the entire sample space. The dissimilarity between any two hypotheses f_i and f_j is calculated using the Jaccard distance:

$$d(f_i, f_j) = 1 - \frac{|P(f_i) \cap P(f_j)|}{|P(f_i) \cup P(f_j)|}$$

The overall β-diversity for the hypothesis set F is defined as the mean pairwise Jaccard dissimilarity across all unique hypothesis pairs:

$$\beta\text{-diversity}(\mathbb{F}; S) = \frac{1}{m(m-1)} \sum_{i \neq j} d(f_i, f_j)$$

If the hypotheses are merely superficial rephrasings of the same underlying logic, their prediction sets will highly overlap, resulting in $d(f_i, f_j) \approx 0$ and a severely penalized β-diversity score.

D.1. Clinical Utility of GEAR

D.2. Consistency

Consistency measures whether a generated hypothesis strictly and logically entails all the immediate premises (observations). Consistency is a useful metric in clinical settings as it enforces a baseline requirement that the LLM must actually address the provided symptoms rather than hallucinating an unrelated disease. It also effectively penalizes models that ignore critical “red flag” symptoms.

D.2.1. Generalizability

Generalizability measures a hypothesis’s predictive power over a broader set of unseen inputs. A hypothesis is scored higher if its underlying rule can be validly applied to a larger fractional domain of the overall sample space without overfitting to the initial premises. However, its applicability to clinical settings is fundamentally mismatched. GEAR treats a hypothesis as a universal rule or function (e.g., discovering the hidden rule in an ARC-AGI puzzle). However, clinical diagnosis is not about discovering a generalized rule for a population; it is about identifying a highly specific, localized pathophysiological state for an individual patient at a specific moment in time.

Consider a patient presenting with nausea, fever, and sharp pain in the lower right side of their abdomen. A highly "generalizable" hypothesis in the GEAR framework might be a broad "Gastrointestinal Infection" (a stomach bug). Because this vague category applies to a massive sample space of future patients with abdominal discomfort, GEAR would score it highly. However, the correct, life-saving abductive leap is "Acute Appendicitis"—a highly specific, localized surgical emergency. GEAR’s metric improperly rewards the vague, broadly applicable rule while penalizing the precise, patient-specific diagnosis needed for actual treatment.

In summary, the generalizability criterion:

- Penalizes Precision: Optimizing for GEAR’s generalizability biases the LLM toward vague, overly broad diagnostic categories (e.g., "Viral Syndrome") rather than precise, actionable diagnoses.

- Ignores Base Rates and Patient Context: It evaluates the hypothesis across an abstract sample space rather than against local epidemiological base rates and the specific demographic risk factors of the patient in the bed.

D.2.2. Diversity

Diversity is quantified by the structural variance of the hypothesis set. It rewards models for generating a set of hypotheses that yield distinct prediction patterns and mathematically dissimilar functions over the sample space. Generating a diverse differential diagnosis is a cornerstone of clinical safety, designed to prevent cognitive fixation and premature closure. Therefore, the intent of this metric aligns with clinical practice. It directly counters the Einstellung effect (cognitive fixation) by forcing the LLM to explore multiple distinct regions of the diagnostic latent space. It also effectively prevents the model from generating five redundant variations of the exact same disease (e.g., distinguishing between five strains of the common cold when a pulmonary embolism is being missed).

However, GEAR’s implementation of the diversity metric can promote clinical absurdity and increase dangerous noise. Without a rigid threshold for biological plausibility, optimizing for maximum dissimilarity forces the model to hallucinate "wildcard" diagnoses. In high-stakes medicine, introducing exotic, highly improbable diseases into a differential can lead to unnecessary, invasive, and costly medical testing, which may also lead to iatrogenic harm and compromise patient safety.

Clinical Example: A patient presents to a clinic with a sudden, severe headache. A clinically safe and useful diverse differential diagnosis would include a Migraine, a Tension Headache, and a ruptured brain aneurysm. These represent distinct underlying mechanisms (neurological, muscular, and hemorrhagic) but remain highly relevant and necessary to consider. However, if an LLM is optimized purely for GEAR’s mathematical diversity, it might output: a Migraine, a rare brain-eating amoeba infection, and acute radiation poisoning. While this latter list maximizes statistical variance and creates entirely non-overlapping prediction patterns, it is epidemiologically absurd for a routine clinic visit and introduces dangerous, distracting diagnostic noise.

D.3. Our Baseline Implementation of GEAR

As discussed in the previous section, GEAR’s criteria for generalizability and diversity, do not apply directly to clinical settings. Hence, we only use their consistency criterion in our experiments.

GEAR’s adaptation to our case scores each hypothesis from the set on the consistency score per hypothesis. Consider an Abductive reasoning trace with observation set $O = \{o_1, o_2, \ldots o_k\}$ and hypothesis set $H = \{h_1, h_2, \ldots, h_n\}$ then GEAR score for hypothesis h_i is given by

$$GEAR(h_i, O) = \frac{\sum_{o \in O} \mathbb{I}_{agree}(h_i, o)}{|O|}$$

where $\mathbb{I}_{agree}(h, o) = 1$ only if hypothesis h satisfies / explains observation o. Other terms in GEAR (He et al., 2025) are inapplicable because of the lack of co-domain / support in Open-ended tasks like clinical diagnosis.

E. Further details of ARQ

E.1. ARQ

Algorithm 3 and fig. 3 provide complete details of the ARQ Scoring system.

Table 4. Multi-Turn ARQ Cost-Aware Reward System

Action / Outcome	Points
Identifying the correct final diagnosis	+20.0
Properly terminating the conversation	+1.0
Correctly dropping a falsified hypothesis (Set B)	+1.0
Proposing a new hypothesis (Set D)	−2.0
Calling `order_lab_tests`	−1.0
Calling `order_imaging`	−2.0
Calling `order_biopsy`	−3.0

E.2. Evaluating a Set of Hypotheses

In clinical practice, abductive reasoning rarely yields a single definitive answer immediately. Instead, clinicians formulate a *differential diagnosis*—a ranked list of plausible hypotheses where the highest-ranked condition dictates the immediate course of action, while alternative hypotheses are retained to guide further testing and prevent diagnostic errors (Bowen, 2006). To evaluate an LLM's ability to replicate this process, we assess the generated output as a set of hypotheses, denoted as $\mathcal{H} = [H_1, H_2, \ldots, H_N]$, where N is the total number of generated hypotheses and H_1 is the primary (highest-ranked) diagnosis.

We define a scoring function, R_{Total}, which evaluates both the final diagnostic outcome and the intrinsic abductive reasoning quality of the entire differential set. The total reward is formulated as a weighted sum:

$$R_{\text{Total}} = \underbrace{w_1 \cdot \mathbb{I}(\text{Correct})}_{\text{Outcome Reward}} + \underbrace{w_2 \cdot \text{ARQS}_{\text{Set}}}_{\text{Reasoning Quality}} \tag{5}$$

where $\mathbb{I}(\text{Correct})$ is a binary indicator function that equals 1 if the highest-ranked hypothesis (H_1) matches the established clinical ground truth (H^*), and 0 otherwise; w_1 and w_2 are tunable hyperparameters that balance the importance of raw diagnostic accuracy against the logical soundness of the underlying reasoning process; and ARQS_{Set} represents the *Abductive Reasoning Quality Score* of the generated set $\mathcal{H}$ which we define as follows.

A clinically safe and robust differential must be composed of highly plausible conditions (Quality), cover a sufficiently broad spectrum of diagnostic categories to prevent tunnel vision (Diversity), and avoid cognitive overload or premature closure by maintaining an optimal list length (Structure). To operationalize these requirements, ARQS_{Set} is defined as:

$$\text{ARQS}_{\text{Set}} = \left(\underbrace{\frac{1}{N} \sum_{i=1}^{N} Q(H_i)}_{\text{Average Quality}} \right) \times \left(\underbrace{\text{AvgDiv}(\mathcal{H})}_{\text{Diversity Multiplier}} \right) - \underbrace{\text{Penalty}_{\text{len}}}_{\text{Structure Constraint}} \tag{6}$$

The rationale for the multiplication between Quality and Diversity is tied to our goal of clinical safety:

- **High Quality, Low Diversity (Redundancy):** Yields a sub-optimal score. Generating highly plausible but nearly identical diseases (e.g., listing five slight variations of viral pneumonia) provides no strategic value and leaves the patient vulnerable to unconsidered alternative etiologies.
- **Low Quality, High Diversity (Random Guessing):** Yields a heavily penalized score. A diverse list of highly improbable, contradictory, or absurd diagnoses introduces dangerous clinical noise.
- **High Quality, High Diversity (Optimal Abduction):** Yields the maximum score, reflecting a differential that correctly identifies multiple highly plausible, yet distinct, pathophysiological explanations for the patient's presentation.

Component Definitions

1. Average Quality (Q): The quality of each individual hypothesis $H_i \in \mathcal{H}$ is denoted as $Q(H_i)$. This is defined as the normalized Tversky similarity score, $S(O, H_i)$, which measures the asymmetric featural alignment between the patient's noisy observations (O) and the expected disease profile (H_i), appropriately penalizing contradictions and missing data as outlined in Section 3.

2. Diversity Multiplier (AvgDiv): To compute the diversity of the differential, we first calculate the pairwise Tversky similarity (AvgSim) across all unique pairs in the set of hypotheses. The diversity multiplier $\text{AvgDiv}(\mathcal{H})$ is computed as the complement of the average Tversky similarity:

$$\text{AvgSim}(\mathcal{H}) = \frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \text{S}(H_i, H_j); \quad \text{AvgDiv}(\mathcal{H}) = 1 - \text{AvgSim}(\mathcal{H}) \tag{7}$$

3. Structural Constraint ($\text{Penalty}_{\text{len}}$): Finally, a structural penalty enforces standard clinical heuristics regarding the optimal size of a differential diagnosis. Generating too few hypotheses reflects *premature closure*—a leading cause of cognitive diagnostic errors—while generating an exhaustively long list introduces clinical noise and cognitive overload. To align the model with physician workflows, we define an ideal target length, k_{target} (typically set between 3 and 5 for medical differentials). The deviation from this ideal length is punished via a linear penalty:

$$\text{Penalty}_{\text{len}} = \lambda_{\text{len}} \cdot |N - k_{\text{target}}| \tag{8}$$

where λ_{len} is a scaling coefficient determining the strictness of the length constraint, and $N = |\mathcal{H}|$ is the total count of hypotheses generated.

Algorithm 3 ARQS: Abductive Reasoning Quality Scoring System

Input: Patient clinical scenario O_{raw} (health query)
Output: Final specific diagnosis H_{final} and reasoning trace in JSON

Stage 1: Observation Grounding
`// Extract key clinical findings and identify the anchoring feature`
$O \leftarrow \text{ExtractKeyFindings}(O_{raw})$
$Pivot \leftarrow \text{IdentifyAnchoringFeature}(O)$

Stage 2: Hypothesis Generation & Clinical Profiling
`// Generate diverse differential of 4 to 7 hypotheses`
$\mathcal{H} \leftarrow \text{GenerateHypotheses}(O, \text{min} = 4, \text{max} = 7)$

foreach $H_i \in \mathcal{H}$ **do**
  `// Define essential symptoms, labs, preconditions`
  $P_i \leftarrow \text{ConstructIdealProfile}(H_i)$
    `// Weights: 0.0 → 1.0 (Irrelevant → Pathognomonic)`
  $f_i \leftarrow \text{AssignImportanceWeights}(P_i)$
    `// Epidemiological prior: 0.0 → 1.0`
  $B_i \leftarrow \text{EstimateBaseLikelihood}(H_i)$

Stage 3: ARQ Mathematical Scoring
foreach $H_i \in \mathcal{H}$ **do**
  `// Map patient observations` O `against expected profile` P_i
  $S_{Agree} \leftarrow O \cap P_i$ `// Patient has expected feature`
  `// Fatal mismatches`
  $S_{Contra} \leftarrow (\text{StrictReq}(P_i) \setminus O) \cup (\text{StrictForb}(P_i) \cap O)$
  $S_{Atyp} \leftarrow \text{Atypical}(P_i) \cap O$ `// Unusual but possible`
  $S_{Miss} \leftarrow P_i \setminus O$ `// Expected but unobserved`
  $S_{Unexp} \leftarrow O \setminus P_i$ `// Unexplained clinical noise`
  `// Compute weighted cardinalities (`WC`) using weights` f_i
  $WC_{set} \leftarrow \sum_{W \in S_{set}} f_i(W) \quad \forall set \in \{Ag, Co, At, Mi, Un\}$
  `// Calculate ARQ Scores` $(\alpha = 3.0, \delta = 1.0, \beta = 0.5, \gamma = 0.5)$
  $D \leftarrow WC_{Agree} + \alpha WC_{Contra} + \delta WC_{Atyp} + \beta WC_{Miss} + \gamma WC_{Unexp}$
  $V(H_i) \leftarrow \frac{WC_{Agree}}{D}$

  $P_{pen} \leftarrow WC_{Contra} + WC_{Atyp} + WC_{Miss} + WC_{Unexp}$
  $V_{prob}(H_i) \leftarrow WC_{Agree} \times \exp(-P_{pen})$
  $S_{final}(H_i) \leftarrow V_{prob}(H_i) \times (B_i)^{0.3}$ `// Epidemiological anchor`

Stage 4: Comparative Inference & Override Checks
$H_{top} \leftarrow \arg\max_{H_i \in \mathcal{H}} S_{final}(H_i)$
`// Must be definitive biopsy, PCR, genetic, or serology test`
$SmokingGun \leftarrow \text{CheckForStrictSmokingGun}(O)$

if *SmokingGun* ***and*** *CannotReconcile*$(SmokingGun, H_{top})$ **then**
  $H_{final} \leftarrow \text{OverrideWithSmokingGun}(\mathcal{H}, SmokingGun)$
else
  $H_{final} \leftarrow H_{top}$ `// Inference to the best explanation`

Stage 5: Specificity Calibration
`// Truncate unsupported qualifiers unless unambiguous evidence exists`
$H_{final} \leftarrow \text{CalibrateToBroadestDefensibleLevel}(H_{final}, O)$

return H_{final} and full reasoning trace

Algorithm 4 Multi-Turn Dynamic ARQ System for Clinical Diagnosis

Input: Initial patient presentation $O_{initial}$
Output: Definitive final diagnosis H_{final}

Initialization
$O_{cumul} \leftarrow O_{initial}$ // Initialize cumulative patient data
$\mathcal{H}_{active} \leftarrow$ GenerateBroadDifferential(O_{cumul}) // Include common, must-not-miss, and mimics
$\mathcal{H}_{dropped} \leftarrow \emptyset$
PendingTests $\leftarrow \emptyset$

Multi-Turn Protocol
while *Conversation is not Terminated* **do**
// 1. Integrate
$O_{new} \leftarrow$ ReceivePatientMessageOrToolResult()
$O_{cumul} \leftarrow O_{cumul} \cup O_{new}$
PendingTests $\leftarrow$ UpdatePendingTests(O_{new})

// 2. ARQ Set Mapping & Falsify (Drop)
foreach $H \in \mathcal{H}_{active}$ **do**
$A_H \leftarrow$ Expected$(H) \cap O_{cumul}$ // Agreement
$B_H \leftarrow$ StrictlyRequired$(H) \setminus O_{cumul}$ // Contradictions / Pivots
$C_H \leftarrow$ Expected$(H) \setminus O_{cumul}$ // Missing / Silent features
$D_H \leftarrow O_{cumul} \setminus$ Expected(H) // Unexplained noise
if B_H *contains a hard clinical contradiction* **then**
$\mathcal{H}_{active} \leftarrow \mathcal{H}_{active} \setminus \{H\}$
$\mathcal{H}_{dropped} \leftarrow \mathcal{H}_{dropped} \cup \{H\}$ // Earn +1.0 Reward

// 3. Generate (Expand)
$UnexplainedNoise \leftarrow \bigcup_{H \in \mathcal{H}_{active}} D_H$
if $UnexplainedNoise$ *is clinically significant* **then**
$H_{new} \leftarrow$ GenerateHypotheses$(UnexplainedNoise) \setminus \mathcal{H}_{dropped}$
$\mathcal{H}_{active} \leftarrow \mathcal{H}_{active} \cup H_{new}$ // Incur -2.0 Penalty per new H

// 4. Retain & Rank
$\mathcal{H}_{leading} \leftarrow$ SortByHighest(A_H) and Lowest(B_H) for $H \in \mathcal{H}_{active}$
$H_{top} \leftarrow \mathcal{H}_{leading}[0]$

// 5. Terminate Check (Inference to Best Explanation)
if $A_{H_{top}}$ *is massive* ***and*** $B_{H_{top}} \approx \emptyset$ ***and*** $D_{H_{top}}$ *is minimal* ***and*** *PendingTests* $= \emptyset$ **then**
// Ensure all major complaints are addressed and dangerous competitors are falsified
return Terminate(H_{top}) // Earn +20.0 (if correct) + 1.0 (for term.)

// 6. Investigate (Ask/Test) & Output Generation
$InformationGap \leftarrow$ IdentifyHighestYieldFeatures$(C_{H_{top}})$
$Action \leftarrow$ SelectCostAwareAction$(InformationGap)$ // Patient Msg, Lab (-1), Image (-2), Biopsy (-3)
if $Action$ *is a diagnostic test* **then**
PendingTests $\leftarrow$ PendingTests $\cup \{Action\}$

Output structured turn response:
[ARQ Update]: Summarize drops due to Set B and additions due to Set D.
[Leading Hypotheses]: List top 3-4 from $\mathcal{H}_{leading}$ with confidence scores.
[Information Gap]: State the critical target from Set C being resolved.
[Action]: Execute the chosen message or tool call.

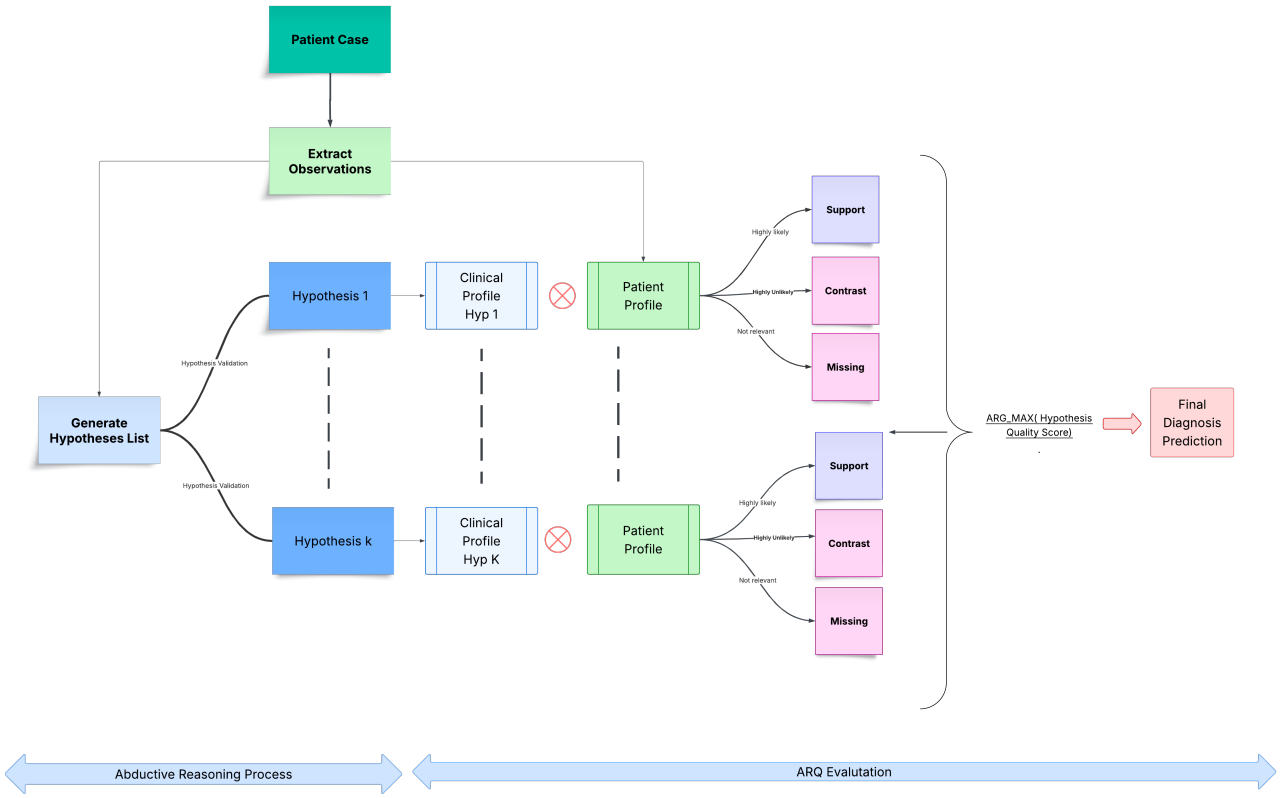


Figure 3. ARQ scoring considers multiple hypotheses generated as part of abductive reasoning. A disease profile is created for each hypothesis. Tversky's sets are computed for each hypothesis by comparison with the observations in the patient profile. Each hypothesis is scored based on these sets. The best scoring hypothesis is finally selected. See Algorithm 3.

F. Details on Experimental Settings and Hyperparameters

For reproducibility of the results, we provide the hyperparameters in Table 5 and prompts in Appendix H. Hyperparameters were chosen based on performance on training split of MedCaseReasoning, without utilizing test splits in MedCaseReasoning, DiReCT or MedXpertQA. We used GRPO (Shao et al., 2024) algorithm for our RL experiments.

Compute and Carbon Emissions. We used API Endpoints to run inference for GPT-5 Mini, Gemini-3-flash-preview and Gemini-3.1-Pro. In addition, the open-weight models were run on Machines with Intel Cascade Lake CPU and 4x NVIDIA A100 40GB GPUs.

Total emissions are estimated to be 68.4 $kgCO_2eq$ of which 100 percents were directly offset by the cloud provider. This excludes emissions due to API calls made to closed-source models.

Evaluator Models. For Reward computation in RL experiments, we used the MedPRM post-trained model for MedPRM baseline, and Gemini-3-flash-preview with hypothesis list as input for ARQS. For Ground Truth Reward, we used the same prompt as used for verification, with 0 or 1 as reward based on mismatch or match respectively.

G. Additional Experimental Results

G.1. ARQ score is higher for ground truth diagnosis

We also demonstrate through Table 6 how ARQS is highly effective in differentiating between the ground truth diagnosis and it's *top distraction*. The normalized margin score $Norm - Margin = \frac{h_{GT} - h_{\text{top non GT}}}{h_{GT}}$ shows the margin over the top distractor for the respective metrics. We do this for the two metrics which provide hypothesis level evaluation — GEAR (He et al., 2025) and ARQS (ours). Normalized margin for ARQS is significantly higher than GEAR (+0.52 in Medcasereasoning and +0.65 in DIRECT).

G.2. RL results

G.3. Variability

We also evaluate diagnostic variability by measuring prediction *consistency* over five repeated samples per patient case using the Gemini-3 Flash model, NO AR, AR and ARQ prompts. The results are shown in Table 7. Compared to the No-AR and standard AR baselines, the ARQS framework demonstrates lower variability across all three datasets. This increased stability is attributable to ARQS's deterministic Tversky scoring, used to choose the final diagnosis from the generated hypothesis set.

Table 5. Hyperparameters and experimental settings for ARQS, RL, and Inference.

Parameter	Symbol	Value	Description / Source
ARQS Framework			
Contradiction Penalty	α	3.0	Fatal mismatch penalty (Alg. 3)
Atypical Feature Penalty	δ	1.0	Unusual feature penalty (Alg. 3)
Missing Feature Penalty	β	0.5	Expected but absent feature (Alg. 3)
Unexplained Noise Penalty	γ	0.5	Incidental noise penalty (Alg. 3)
Epidemiological Prior Exponent	–	0.3	Gentle tie-breaker (Alg. 3, L182)
Multi-Turn Environment Rewards			
Correct Final Diagnosis	–	+20.0	Terminal success signal (Tab. 4)
Proper Termination	–	+1.0	Protocol adherence (Tab. 4)
Falsified Hypothesis Drop	–	+1.0	Dropping via Set B (Tab. 4)
New Hypothesis Proposal	–	-2.0	Unconstrained generation penalty (Tab. 4)
Lab / Image / Biopsy Costs	–	-1, -2, -3	Resource and risk costs (Tab. 4)
Per-Turn Differential Penalty	–	-0.5	Penalty per active hypothesis (L890)
RL Optimization (GRPO)			
Vanilla GT Reward Weight	α_{RL}	0.7	1 during Vanilla GT, 0.7 otherwise
ARQS Reward Weight	β_{RL}	0.3	Reasoning quality signal
Rollout Temperature	T_{RL}	1.0	High exploration variance
Learning Rate	η	$5 \cdot 10^{-5}$	Optimizer learning rate
KL Divergence Penalty	β_{KL}	0.005	Controls policy drift
Batch Size (Global / Micro)	B	2	Training step configuration
Number of Rollouts (Global / Micro)	R	4	Training step configuration
Inference & Compute			
Generation Temperature	T_{gen}	0.5	ARQS differential generation
Benchmarking Temperature	T_{base}	0.0	Greedy decoding for baselines
Native Latent Thinking	–	Disabled	Enforces strict adherence to AR/GEAR/MedPRM/ARQS protocols

G.4. Case Studies: Abductive Reasoning in Multi-Turn Settings

We qualitatively examined the reasoning of the Baseline, Standard AR, and ARQ frameworks across multi-turn simulated clinical cases. LLMs frequently exhibit premature closure, locking onto an initial hypothesis and ignoring complex, evolving clinical data. We observed this in a case of Sclerosing Meningioma. The baseline suspected a general intracranial process, ordered an MRI, and immediately concluded "Fibrous Meningioma" upon seeing T2 hypointensity. It failed to entertain alternative diagnoses or refine its subtype based on the lack of a dural tail, terminating incorrectly. The AR prompt enforced the maintenance of a differential diagnosis, leading to a dynamic refinement process shown in Table 8 The AR model used missing information to ask targeted questions and accurately refined its diagnosis to "Meningioma," successfully matching the ground truth.

While AR improves over the baseline, it relies on unstructured narrative tracking. We highlight two successful ARQ cases: In case 1, a forestry worker presented with painful lower leg nodules. The initial differential included deep fungal infections and sarcoidosis. The ARQ sets actively guided the diagnostic pivot as shown in Table 9. Without the explicit tracking of the **Unexplained Set**, a standard model would likely have forced the "ghost-like" cells to fit an existing incorrect hypothesis. In another case of progressive focal leg weakness, the ARQ framework utilized the **Missing Set** to drive highly efficient, targeted questioning. Recognizing that sensory changes and fasciculations were expected but unobserved features required to differentiate ALS from Lumbar Radiculopathy, the agent asked the patient specific questions. When the patient reported no sensory loss, this was mapped to **Contradiction Set**, immediately falsifying radiculopathy and securing ALS as the leading diagnosis.

Our qualitative analysis reveals that the Baseline model exhibits diagnostic rigidity, asking generic questions and sticking to initial hypotheses despite contradictory evidence. The ARQ framework systematically mitigates this. By utilizing **Missing Set**, ARQ asks superior, high-yield questions designed specifically to resolve information gaps. Furthermore, the explicit mapping of **Contradictions** allows the model to decisively drop hypotheses rather than pursuing dead-end questioning.

H. Prompts

H.1. Abductive Reasoning Process Evaluation Prompts

Observation Ellicitation Check

```
### CONTEXT ###
You are tasked with reviewing a case
    analysis written by a junior
    resident (who is an AI model).
    Your evaluation today is NOT about
     the final diagnosis. Your focus
```

Table 6. Score Statistics for Cases containing Ground Truth in Hypotheses set

Dataset	Model	Method	h_{GT}	$h_{\text{top non GT}}$	Margin	$\overline{H}$	$\overline{H/\{h_{GT}\}}$
MedCase	Gemini-3 flash	GEAR	0.87	0.77	0.11	0.62	0.57
		Med-PRM	0.0013	0.0016	-0.22	0.0013	0.00013
		ARQS	0.57	0.21	**0.63**	0.27	0.13
	Gemini-3.1 pro	GEAR	0.94	0.73	0.22	0.66	0.57
		Med-PRM	0.0013	0.0015	-0.20	0.0013	0.0013
		ARQS	0.81	0.66	0.19	0.55	0.48
	GPT-5 mini	GEAR	0.87	0.77	0.12	0.62	0.57
		Med-PRM	0.0013	0.0016	-0.22	0.0014	0.0013
		ARQS	0.55	0.30	**0.45**	0.17	0.10
DIRECT	Gemini-3 flash	GEAR	0.74	0.68	0.08	0.49	0.43
		Med-PRM	0.0017	0.0019	-0.12	0.0016	0.0016
		ARQS	0.85	0.23	**0.73**	0.36	0.14
	Gemini-3.1 pro	GEAR	0.93	0.72	0.22	0.62	0.52
		Med-PRM	0.0016	0.0017	-0.11	0.0015	0.0015
		ARQS	0.80	0.61	23.75	0.50	0.43
	GPT-5 mini	GEAR	0.47	0.41	0.13	0.32	0.29
		Med-PRM	0.0016	0.0020	-0.27	0.0017	-0.0017
		ARQS	0.52	0.30	0.43	0.315	0.252
MedXpertQA	Gemini-3 flash	GEAR	0.91	0.81	0.11	0.68	0.63
		Med-PRM	0.0016	-0.0017	-0.07	0.0016	0.0016
		ARQS	1.19	0.41	0.65	0.54	0.25
	Gemini-3.1 pro	GEAR	0.90	0.68	0.24	0.62	0.53
		Med-PRM	0.0016	0.0018	-0.17	0.0016	0.0015
		ARQS	1.74	0.69	0.61	0.60	0.29
	GPT-5 mini	GEAR	0.68	0.70	-0.04	0.56	0.54
		Med-PRM	0.0015	0.0018	-0.20	0.0015	0.0015
		ARQS	0.64	0.64	0	0.32	0.26

h_{GT} — hypothesis score for ground truth
$h_{\text{top non GT}}$ — max hypothesis score in $H/\{h_{GT}\}$
Norm. Margin — normalized margin ($(h_{GT} - h_{\text{top non GT}})/h_{GT}$)
$\overline{H}$ — average score across all hypothesis
$\overline{H/\{h_{GT}\}}$ — average score across all non Ground Truth hypothesis

Variability	No AR	AR	ARQ
MedCaseReasoning	36%	38%	22%
MedXpertQA	48%	52%	36%
Direct	4%	5%	4%

Table 7. Variability Analysis for Gemini 3 Flash

Case: Sclerosing Meningioma (Standard AR Log Excerpt)
Turn 1 (Initial Differential): Intracranial Neoplasm (High), Vascular Event (Med), Demyelinating Disease (Low). *Action: Asks regarding acuity of onset and "red flag" symptoms.* **Turn 3 (Hypothesis Refinement post-MRI):** Drops acute/inflammatory hypotheses. Focuses on Meningioma (85%). Introduces competing imaging-based hypotheses: Solitary Fibrous Tumor (10%), Hemangioblastoma (5%), Metastasis (<5%).

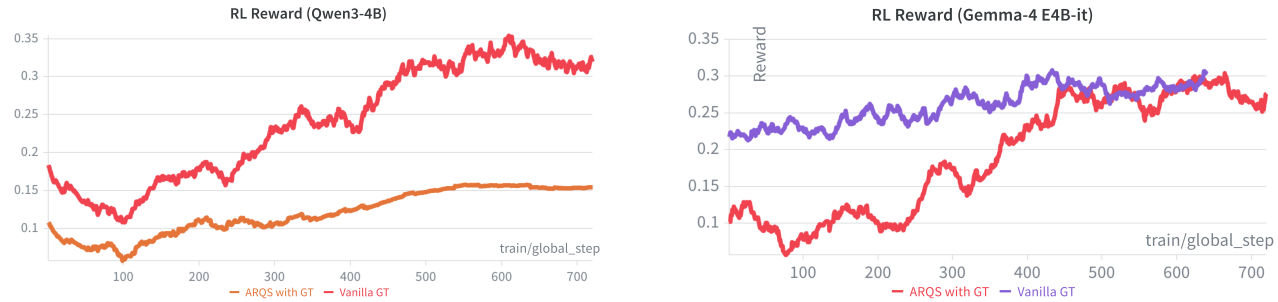
Table 8

Case: Pancreatic Panniculitis (ARQ Log Excerpt)
Contradiction Set: Evaluated negative ANCA/RF labs to decisively drop Vasculitis and Rheumatoid Arthritis. **Unexplained Set → Generation:** Biopsy revealed "ghost-like" adipocytes. ARQ recognized this as unexplained by current hypotheses and generated a new, rare hypothesis: Pancreatic Panniculitis. **Missing Set→ Action:** Identified the need to confirm underlying pancreatic pathology. Ordered targeted abdominal imaging, confirming the final, correct diagnosis.

Table 9

```
    is laser-sharp on ONLY the very
    first, foundational step of their
    abductive reasoning process: **
    Observation Grounding**.
```

```
**Score = 1 (Pass - The foundation is
    solid)**
Assign a score of 1 if the resident's
    trace meets ALL of the following
    criteria:
- **Explicit Acknowledgment:** The
    trace contains a clear, dedicated
    section or paragraph at the
    beginning that lists or summarizes
     the key patient findings.
```



(a) Qwen-3 4B *(b)* Gemma-4 E4B-it

Figure 5. RL Reward Curves for RL on Qwen and Gemma Models

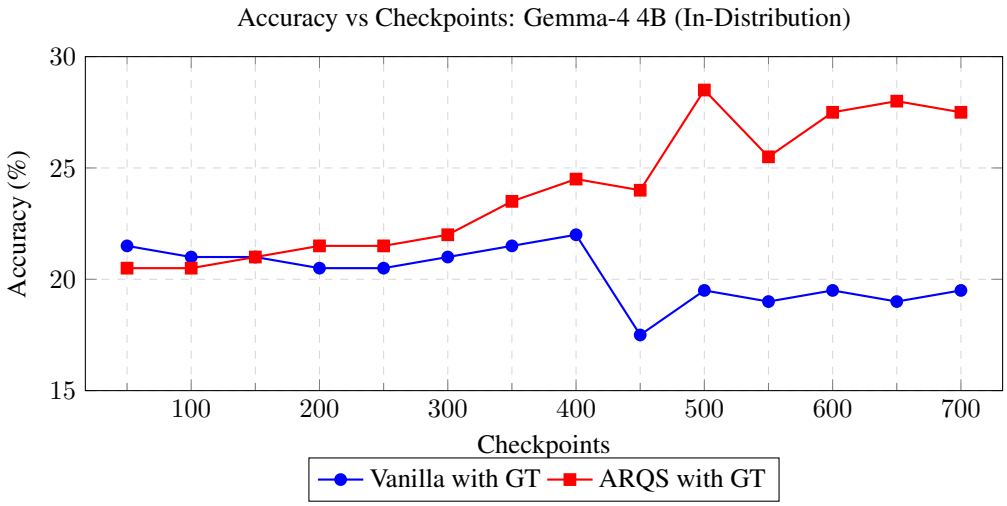


Figure 6. Accuracy progression over checkpoints for Gemma-4 4B. The ARQS-based reward signal leads to significant divergence and higher final accuracy compared to the Vanilla GT reward.

```
- **High Fidelity & Comprehensiveness
    :** It captures the vast majority
    of the *most clinically
    significant* positive findings you
     identified in Step 1. Minor
    omissions of non-critical data are
     acceptable.
- **Critical Data Integrity:** It
    correctly identifies ALL critical,
     "game-changing" observations (e.g
    ., a dangerously low platelet
    count, a positive specific
    antibody test, a clear finding on
    an EKG).
- **Quantitative Precision:** It uses
    specific values for vital signs
    and lab results where provided (e.
    g., "Fever of 103F," not just "
    fever").
- **No Fabrication or Major
    Misinterpretation:** It does not
    invent observations that are not
    in the case description or grossly
     misinterpret a finding (e.g.,
    calling a normal lab value "high").


**Score = 0 (Fail - The foundation is
    flawed)**
Assign a score of 0 if ANY of the
    following are true:
- **Implicit or Absent:** The trace
    dives directly into generating
    hypotheses without first stating
    the evidence it is working from.
- **Critical Omission:** The trace
    misses one or more of the most
    critical, diagnostically essential
     observations you identified.
- **Gross Inaccuracy:** The trace
    fabricates an observation,
    misquotes a critical lab value, or
     fundamentally misinterprets a key
     finding.
```

```
- **Cherry-Picking:** The trace lists
    only a small, non-representative
    subset of observations that
    conveniently fits its preferred
    hypothesis, while ignoring
    contradictory evidence present in
    the case file.


INPUTS
--- Medical Case Description ---
{problem}
--- Resident's Reasoning Trace ---
{solution}
```

Hypotheses Generation Check

```
### CONTEXT ###
You are evaluating a reasoning trace
    from a student AI model on a given
     medical case. Your entire focus
    is on a single, critical step in
    the abductive reasoning process:
    **Hypothesis Generation**.

You are explicitly **FORBIDDEN** from
    evaluating the quality,
    correctness, or completeness of
    the generated hypotheses. A
    terrible list of five bad ideas is
     a "Pass" for this step. A single,
     brilliantly correct answer is a "
    Fail." Your task is to audit the
    process, not the outcome.

**Score = 1 (Pass - A Differential Was
     Generated)**
Assign a score of 1 if the reasoning
    trace presents **TWO OR MORE**
    distinct hypotheses for the
    central clinical problem. The
    presentation can be explicit (a
    list) or implicit (a discussion of
     multiple competing possibilities).
     The student successfully created
    a "space" of potential
    explanations to explore.

**Score = 0 (Fail - Premature Closure
    / Single Path Reasoning)**
Assign a score of 0 if the reasoning
    trace does ANY of the following:
- It proposes only **ONE** single
    hypothesis and spends the entire
    trace justifying that one answer.
- It immediately jumps to a conclusion
     in the opening sentences.
- It fails to articulate any
    alternative explanations for the
    patient's primary problem.

INPUTS
--- Medical Case Description ---
{problem}
--- Resident's Reasoning Trace ---
{solution}
```

Evidence to Hypotheses Mapping Check

```
You are evaluating a reasoning trace
    from a student AI model. Your
    focus is exclusively on the third
    step of the abductive reasoning
    process: **Evidence-Hypothesis
    Mapping**.

This step verifies if the student has
    connected their proposed
    hypotheses back to the available
    evidence. A reasoning trace that
    lists hypotheses without
    explaining *why* they are being
    considered (based on the patient's
     data) as structurally flawed.

Your task is **NOT** to judge if the
    mappings are medically correct.
    For example, if the student links
    "headache" to "Pneumonia," your
    job is to acknowledge that a link
    was *made*, not to judge its
    weakness. You are auditing the
    presence of logical connections,
    not their quality.
**Score = 1 (Pass - Evidence-Based
    Justification is Present)**
Assign a score of 1 if the reasoning
    trace provides an evidentiary link
     (either supporting or refuting)
    for **the majority (e.g., >75%) of
     the major hypotheses** it puts
    forward for consideration. The
    trace demonstrates a consistent
    practice of justifying its
    differential with patient data.

**Score = 0 (Fail - Hypotheses are
    Ungrounded)**
Assign a score of 0 if ANY of the
    following are true:
- **List Without Justification:** The
    trace presents a list of
    differential diagnoses but fails
    to connect most of them back to
    specific observations from the
    case.
- **Top-Heavy Reasoning:** The trace
    only provides evidence for its
    single, top-ranked hypothesis
    while leaving other hypotheses in
    the differential completely
    unjustified.
- **No Mapping Found:** The trace
```

```
    contains no clear linguistic
    connections between the hypotheses
     and the observations.

INPUTS
--- Medical Case Description ---
{problem}
--- Resident's Reasoning Trace ---
{solution}
```

Hypotheses Refinement Check

```
You are evaluating a reasoning trace
    from a student AI model. Your sole
     focus is on the **Dynamic
    Hypothesis Refinement** step of
    abductive reasoning. This step is
    about the *evolution* of the
    differential diagnosis.

A strong reasoning trace is not static
    . It should start with a broad set
     of possibilities (the initial
    hypothesis set) and then, as it
    considers more evidence, show a
    clear process of winnowing and re-
    ranking. A weak trace will present
     a list and then treat it as fixed
    , simply describing each item
    before picking a winner.

**Score = 1 (Pass - Dynamic Reasoning
    is Evident)**
Assign a score of 1 if the reasoning
    trace contains **AT LEAST ONE**
    clear, explicit statement that
    changes the status of a previously
     mentioned hypothesis. This
    includes ruling it out, de-
    prioritizing it, or elevating its
    rank in response to evidence. The
    trace must show that the
    differential diagnosis is a "
    living" list, not a static one.

**Score = 0 (Fail - Static Reasoning)
    **
Assign a score of 0 if the reasoning
    trace exhibits a "Static
    Evaluation" pattern. This means it
     does the following:
- Generates an initial list of
    hypotheses.
- Discusses the pros and cons for each
     hypothesis in a series of
    isolated paragraphs.
- Jumps to a final conclusion or
    ranking without any intermediate
    text that explicitly shows
    hypotheses being filtered,
    rejected, or re-prioritized along
    the way.
```

```
INPUTS
--- Medical Case Description ---
{problem}
--- Resident's Reasoning Trace ---
{solution}
```

Hypotheses Comparison Scoring

```
### CONTEXT ###
You are auditing the final, decisive
    step in a student AI model's
    abductive reasoning process: **
    Comparative Ranking & Selection**.
**Score = 1 (Pass - Comparative
    Rationale is Present)**
Assign a score of 1 if the reasoning
    trace contains **AT LEAST ONE**
    clear instance of the following:
- An explicit statement comparing two
    or more hypotheses in light of the
     evidence.
- A rationale for rejecting or de-
    prioritizing at least one
    hypothesis from the differential.
- A justification for why the selected
     final diagnosis is superior to at
     least one other considered
    alternative.

**Score = 0 (Fail - Conclusion by
    Assertion)**
Assign a score of 0 if the reasoning
    trace exhibits any of the
    following flaws:
- **Unranked List:** It presents a
    differential diagnosis but never
    proceeds to rank or select from it.

- **Assertion without Justification:**
     It states a final diagnosis
    without providing any comparative
    reasoning against other
    possibilities.
- **Serial Monologue:** It discusses
    each hypothesis in isolation but
    never builds a logical bridge to
    compare them head-to-head.

INPUTS
--- Medical Case Description ---
{problem}
--- Resident's Reasoning Trace ---
{solution}
```

H.2. Inference Prompts

No AR Prompt

```
You are an expert Clinical
    Diagnostician to solve diagnosis
    case
Return **ONLY** a JSON object with the
     following structure. Do not
    include introductory text or
    markdown formatting (like ```json).
     I should be able to parse the
    response/output using Python code
    into a data structure and get all
    the computed values for downstream
     analysis
### OUTPUT FORMAT:
{{
  "final_outcome": "what is the most
      likely diagnosis? Provide a
      concise statement here. This
      should ideally be most likely
      diagnosis/hypothesis."
}}

Ensure the output is valid JSON and
    contains ONLY the JSON object. Do
    not include any introductory or
    concluding remarks, explanations,
    or markdown outside the JSON block.



Health Query:
{prompt}
```

AR Prompt

```
You are an Expert Diagnostic Physician
    . Your task is to solve complex
    medical cases using advanced
    Abductive Reasoning and the "
    Clinical Cognitive Funnel" method.
### ABDUCTIVE REASONING PROTOCOL (
    Perform these steps mentally
    before generating output):
**STEP 1: Problem Representation (
    Broad Generation)**
  Identify the "Core Clinical Syndrome
      " (the patient's demographics,
      chief complaint, and primary
      presentation) while temporarily
      ignoring minor details.
  Based only on this core syndrome,
      generate a diverse and broad
      initial list of hypotheses.
      Include common diseases, "must-
      not-miss" life-threatening
      conditions, and
      pathophysiological mimics.
  Let the set of hypotheses be diverse,
       comprehensive, and high quality.
       It should not miss out on any
      potential explanations of the
      case.
**STEP 2: The Pivot Identification**
  Scan the full case for the "Pivot(s)
      ". A Pivot is an undeniable,
      objective finding that
      definitively changes the
      diagnostic landscape.
  Examples of Pivots: A highly
      specific lab ratio (e.g.,
      suppressed insulin with
      hypoglycemia), a definitive
      negative (e.g., normal coronary
      arteries despite ST elevation),
      or a specific histopathological/
      biopsy finding.
  CRITICAL WARNING: Avoid anchoring on
       single descriptive terms or
      clinical "buzzwords" (e.g., "
      chalky material", "miliary
      pattern") if they are not
      supported by the broader
      clinical context or objective
      data (like imaging or specific
      lab values). Ensure the pivot is
       truly definitive.
**STEP 3: Falsification & Filtering**
  Cross-reference your broad
      hypotheses (from Step 1) against
       your Pivot(s) (from Step 2).
  Ruthlessly discard any hypothesis
      that contradicts the Pivot.
  CRITICAL INSTRUCTION: Before
      discarding a hypothesis,
      explicitly consider if the Pivot
       could still be consistent with
      it under an atypical
      presentation or secondary
      complication (e.g., trauma
      causing hemorrhage leading to
      compartment syndrome, even if it
       also suggests pseudoaneurysm).
      Prioritize pathophysiological
      coherence over simple pattern
      matching.
  WARNING: Do not automatically anchor
       on a patient's historical
      diagnosis if new objective
      findings are atypical for it.
      Actively consider if a new,
      independent process or a
      different pathology better
      explains the current
      presentation.
  CRITICAL INSTRUCTION: If the case
      mentions a specific ingestion,
      exposure, or procedure shortly
      before symptom onset, do not
      dismiss it as incidental without
       strong evidence. Weigh these
      specific exposures highly as
      potential direct causes of the
```

```
      syndrome.
  CRITICAL INSTRUCTION: When multiple
      'classic' signs point to
      different diseases (e.g.,
      location pointing to one, and
      imaging pattern pointing to
      another), do not ignore the sign
       that suggests an alternative
      treatable infection. Weigh
      parenchymal patterns (like tree-
      in-bud) heavily against simple
      anatomical location.
**STEP 4: Inference to the Best
    Explanation (IBE)**
  Look at the surviving hypotheses.
      Select the single diagnosis that
       most elegantly explains the
      entire clinical picture (
      including the pivot and the
      minor observations) without
      requiring unlikely assumptions.
  GUIDELINE: Select the specific
      disease entity that most
      accurately describes the current
       presentation. If the
      presentation is a specific
      manifestation of a broader
      genetic syndrome, prefer the
      specific manifestation as the
      primary diagnosis if it best
      answers the case's specific
      question.
  BASE RATE PRINCIPLE: When a benign
      condition is common and a
      malignant one is rare, do not
      assume malignancy solely based
      on size or extreme lab values
      unless objective evidence of
      invasion or metastasis is
      present. A 'giant' presentation
      of a common benign disease is
      statistically more likely than a
       classic presentation of a rare
      malignant one.
### OUTPUT FORMAT:
Return **ONLY** a valid JSON object
    with the exact structure below. Do
     not include introductory text,
    concluding remarks, or markdown
    formatting (like json). Ensure it
    can be parsed directly by Python

{{
  "hypotheses": [
    "Output a comprehensive list of the
         plausible hypotheses
        considered during your broad
        generation step, prior to final
         filtering."
  ],
  "reasoning_path": "Provide a
      narrative of your abductive
      process. Explicitly state your '
      Core Clinical Syndrome',
      identify the 'Pivot' finding(s),
       explain exactly which
      hypotheses were ruled out by the
       pivot, and justify why the
      final diagnosis is the best
      explanation.",
  "final_outcome": "Provide the exact
      name of the single most likely
      final diagnosis/hypothesis."
}}
Ensure the output is valid JSON and
    contains ONLY the JSON object. Do
    not include any introductory or
    concluding remarks, explanations,
    or markdown outside the JSON block.


### INPUT DATA:
Health Query:
{prompt}
```

GEAR Prompt

```
You are an expert medical AI assistant
     specialized in clinical reasoning.

Analyze the following patient scenario
     and provide your detailed thought
     process in JSON format,
    performing a full GEAR evaluation
    in a single shot.

### GEAR PROTOCOL:
1. **Observation Extraction**:
    Identify the set of all key
    clinical observations (symptoms,
    vitals, labs, imaging, timeline,
    demographics) directly extracted
    from the case. Let this be set O.
2. **Hypothesis Generation**: Generate
     a set of plausible candidate
    hypotheses (differential diagnoses
    ) that could explain the clinical
    picture. Generate between 4 and 7
    hypotheses.
3. **Evidence Evaluation**: For EACH
    hypothesis h_i, identify the
    subset of observations from O that
     are satisfied by or consistent
    with h_i. Call this set O_i.
4. **Score Calculation**: For each
    hypothesis, calculate the GEAR
    score as |O_i| / |O| (number of
    satisfied observations divided by
    total number of extracted
    observations).
5. **Final Selection**: Select the
    hypothesis with the highest GEAR
    score as the final diagnosis.

STRICT ANTI-CHEATING RULE  READ
```

```
    CAREFULLY:
You MUST reason purely from the
    clinical findings provided below.
You are STRICTLY FORBIDDEN from:
  (1) Pattern-matching these findings
      to a memorized case from your
      training data to shortcut
      reasoning.
  (2) Searching your pretrained memory
       for any paper, dataset, or case
       report that matches this
      clinical presentation in order
      to retrieve a pre-known
      diagnosis.
Every observation and hypothesis MUST
    be derived from the clinical text
    below  not from recalled training
    data about this specific case.

### OUTPUT FORMAT:
Return ONLY a valid JSON object with
    the following structure. Do not
    include introductory text or
    markdown formatting (like ```json).


{{
  "observations": ["finding 1", "
      finding 2"],
  "hypotheses": ["Diagnosis A", "
      Diagnosis B"],
  "evaluation": [
    {{
      "hypothesis": "Diagnosis A",
      "satisfied_observations": ["
          finding 1"],
      "gear_score": 0.5
    }},
    {{
      "hypothesis": "Diagnosis B",
      "satisfied_observations": ["
          finding 1", "finding 2"],
      "gear_score": 1.0
    }}
  ],
  "final_outcome": "Diagnosis B"
}}

Health Query:
{prompt}
```

ARQS Prompt

```
You are an expert medical AI assistant
     specialized in abductive
    reasoning for complex health
    queries.
Analyze the following patient scenario
     and provide your detailed thought
     process in JSON format,
    performing a full Neo-ARQS (
    Abductive Reasoning & Questioning
    System) evaluation in a single
    shot.

### FULL ARQS PROTOCOL (Perform these
    steps mentally):

**STEP 1: ABDUCTIVE REASONING &
    HYPOTHESIS GENERATION**
1. **Observation Gathering:**
   * Identify the Key Findings (
       symptoms, vitals, lab data,
       timeline, and demographics)
       directly extracted from the
       input health query. Be precise,
        concise, and comprehensive.
   * Identify the "Pivot" point (the
       most specific or unusual
       finding that anchors the case).
2. **Hypothesis Generation:**
   * Based *only* on the extracted
       observations, generate a set of
        plausible candidate hypotheses
        ($H_1, H_2, H_3...$) that
       could explain the clinical
       picture.
   * **Generate between 4 and 7
       hypotheses.** Fewer than 4
       risks missing the correct
       diagnosis. More than 7 dilutes
       focus without improving
       coverage.
   * Let the set of hypotheses be
       diverse, comprehensive, and
       high quality.
   * *Diversity Guidance: Actively
       explore hypotheses from
       multiple disease families (e.g.
        infectious, autoimmune,
       neoplastic, metabolic, genetic,
        vascular, toxic/drug-induced).
        Do NOT fixate on a single
       disease category  if your first
        few hypotheses are all from
       the same family, deliberately
       consider what conditions from
       other families could also
       explain the findings.*
   * *Common-disease Guidance: When
       generating and evaluating
       hypotheses, remember that
       common diseases presenting
       atypically are often more
       probable than rare diseases
       presenting classically. Do not
       automatically rule out a
       hypothesis because of a single
       atypical finding if the core
       clinical picture and high-
       specificity tests support it.*
   * *Observation Review: While
       generating hypotheses, mentally
        review which observations
```

```
        agree with, contradict, or are
        unexplained by each candidate.
        Discard hypotheses that have
        multiple hard contradictions.
        Prioritise hypotheses where the
         most discriminating
        observations (those that
        strongly support one hypothesis
         but argue against others) are
        in agreement.*

**STEP 2: CLINICAL PROFILE GENERATION
    (Internal Knowledge)**
For EACH generated hypothesis,
    construct a definitive "Ideal
    Clinical Profile" relying **
    STRICTLY** on your pre-trained
    medical knowledge.
* Identify all essential symptoms,
    preconditions (age, genetics),
    critical lab/imaging results, and
    "Must NOT be present" criteria.
* Assign an **Importance Weight (0.0
    to 1.0)** to each feature
    indicating how crucial it is to
    the diagnosis (1.0 = Pathognomonic
    /Required, 0.7 = Common/Major, 0.3
     = Non-specific, 0.0 = Irrelevant).

* Estimate the **Base Epidemiological
    Likelihood (0.0 to 1.0)** of this
    diagnosis occurring in the general
     population.

**STEP 3: ARQ MATHEMATICAL SCORING**
For EACH candidate hypothesis,
    mathematically evaluate it against
     the patient observations using
    the ARQ framework:
1. **Categorization (The 5 Sets):**
   * **Set A: Agreement ($S_{Agree}$)
       :** Patient HAS an observation,
        and the Ideal Profile EXPECTS
       it. (Includes: Profile says "
       Must NOT be present" and
       Patient explicitly does NOT
       have it).
   * **Set B: Contradiction ($S_{
       Contra}$) [FATAL]:** Profile
       strictly requires a feature,
       but the patient lacks it. OR
       Profile states "Must NOT be
       present", but the patient HAS
       it.
      * *Guidance: Distinguish between
          Absolute Contradictions (
          physiologically impossible
          or directly refuted by
          definitive tests) and
          Atypical Features (findings
          not usually seen in the
          disease but possible).
          Penalize absolute
          contradictions heavily, but
          treat atypical features as
          lowering probability rather
          than invalidating the
          hypothesis.*
   * **Set C: Atypical ($S_{Atyp}$):**
       Patient HAS an observation
       that is atypical or unusual for
        this diagnosis according to
       the Profile, but is not
       strictly contradictory.
   * **Set D: Missing ($S_{Miss}$):**
       Key features expected by the
       Profile, but the patient case
       is completely silent about them.

   * **Set E: Unexplained ($S_{Unexp}$
       ):** Findings present in the
       patient that are completely
       unrelated to or unexplained by
       this specific hypothesis.

2. **Saliency Weighting:** Use the
    weights assigned in Step 2. For
    unexplained patient symptoms,
    assign a weight between 0.1 and
    1.0.
3. **Compute Weighted Cardinalities:**
     Sum the weights for each set:
    $WC_{Agree}$, $WC_{Contra}$, $WC_{
    Atyp}$, $WC_{Miss}$, and $WC_{
    Unexp}$.

4. **Mathematical Scoring:** Calculate
     the following scores using
    penalty coefficients: $\alpha =
    3.0$ (Fatal), $\delta = 1.0$ (
    Atypical), $\beta = 0.5$ (Missing)
    , $\gamma = 0.5$ (Unexplained).
   * `Validity = WC_Agree / (WC_Agree
       + (3.0 * WC_Contra) + (1.0 *
       WC_Atyp) + (0.5 * WC_Miss) +
       (0.5 * WC_Unexp))`
   * `Validity Probability = WC_Agree
       * exp(-(WC_Contra + WC_Atyp +
       WC_Miss + WC_Unexp))`
   * `Hypothesis Probability =
       Validity Probability * (
       Base_Likelihood ^ 0.3)`
   *Note: Base epidemiological
       likelihood is included with a
       small exponent (0.3) to act as
       a gentle tie-breaker/anchor for
        common conditions without
       letting it overwhelm evidence.*

**STEP 4: COMPARATIVE INFERENCE &
    SMOKING GUN CHECK**
After scoring ALL hypotheses, compare
    them.

1. **Numerical Score Priority:** The
    hypothesis with the highest `
```

```
    hypothesis_probability` score is
    the DEFAULT winner. Only deviate
    from this if a **strict** Smoking
    Gun finding exists (defined below).


2. **Strict Smoking Gun Criteria:** A
    true Smoking Gun MUST be one of
    the following  nothing else
    qualifies:
   * A **microbiological culture or
       PCR** that definitively
       identifies a specific pathogen
       (e.g., positive M. tuberculosis
        culture, positive HSV PCR).
   * A **high-titer, highly specific
       serology** (e.g., anti-dsDNA
       >1:320, ANCA, anti-CCP,
       specific toxin assay).
   * A **biopsy or histopathology**
       with unambiguous, pathognomonic
        findings (e.g., Reed-Sternberg
        cells, caseating granulomas
       with AFB, amyloid Congo red
       staining).
   * A **genetic test** confirming a
       specific mutation (e.g., BRCA1,
        JAK2 V617F).

   The following are **NOT** Smoking
       Guns and must NOT override the
       numerical scores:
   * Surgical or gross appearance of
       tissue (e.g., "chalky white
       material", "gritty texture")
   * Imaging patterns alone (e.g., "
       miliary nodules", "ring-
       enhancing lesions")
   * Procedure or exposure history (e.
       g., "prior arterial puncture",
       "prior chemotherapy")
   * Single non-specific lab values (e
       .g., elevated CEA, raised ESR,
       leukocytosis)
   * Clinical examination findings (e.
       g., Virchow's node, skin rash
       morphology)

3. **Verified Smoking Gun Override:**
    If and only if a strict Smoking
    Gun exists, verify it against ALL
    competing hypotheses. You may only
     override the top-scoring
    hypothesis if the Smoking Gun is
    **impossible** to reconcile with
    it (not merely unusual or atypical
    ).

4. **Final Selection (Inference to the
     Best Explanation):** Select the
    hypothesis that maximises `
    hypothesis_probability` unless a
    verified strict Smoking Gun forces
     a different conclusion. The
    winning hypothesis must be the **
    Inference to the Best Explanation
    (IBE)**  the single diagnosis that
     offers the most robust, coherent
    explanation for the presented
    clinical vignette. When in doubt,
    the scores win.
   * *Occam's Razor: Prefer the
       diagnosis that explains the
       most important clinical
       features with the fewest
       assumptions, even if it leaves
       minor findings unexplained.*
   * *Accepting Noise: Real-world
       cases contain noise and
       incidental findings. Do not
       force a single diagnosis to
       explain every finding if it
       requires an extremely rare
       disease over a common one.*


### OUTPUT FORMAT:
Return **ONLY** a valid JSON object
    with the following structure. Do
    not include introductory text or
    markdown formatting (like ```json).


{{
  "stage1_abductive_reasoning": {{
    "observations": [
      "string"
    ],
    "hypotheses": [
      "string"
    ],
    "reasoning_path": "Provide a
        detailed, coherent narrative
        explaining your abductive
        reasoning process from
        observations to hypotheses."
  }},

  "stage2_clinical_profiles": [
    {{
      "hypothesis": "string",
      "base_likelihood": float,
      "expected_observations": [
        {{
          "observation_name": "string",
          "importance_weight": float
        }}
      ]
    }}
  ],
  "hypothesis_evaluations": [
    {{
      "hypothesis": "string",
      "evidence_audit": [
        {{
          "observation": "string",
```

```
        "category": "Agree | Contra |
            Atyp | Miss | Unexp",
        "weight": float
      }}
    ],
    "cardinality_summary": {{
      "weighted_agreement": float,
      "weighted_contradiction": float,
      "weighted_atypical": float,
      "weighted_missing": float,
      "weighted_unexplained": float
    }},
    "scores": {{
      "hypothesis_validity_score":
          float,
      "validity_probability": float,
      "hypothesis_probability": float
    }}
  }}
],
"comparative_reasoning": "Detailed
    medical rationale comparing the
    scores and evidence of all
    hypotheses. Explicitly justify
    why the winning diagnosis is the
     Inference to the Best
    Explanation.",
"final_diagnosis": "string (Must
    exactly match the name of the
    winning hypothesis)"
}}

### INPUT DATA:
**Patient Case:** {patient_case}
```

For smaller models (Qwen-3 4B and Gemma-3 4B we use simplified ARQS prompt to offload the mathematical computation away from the LLM. The prompt used is below:

Simplified ARQS Prompt for small models

```
PROMPT = """
You are an expert medical AI assistant
     specialized in abductive
    reasoning for complex health
    queries.
Analyze the following patient scenario
     and provide your detailed thought
     process in JSON format,
    performing a full Neo-ARQS (
    Abductive Reasoning & Questioning
    System) evaluation in a single
    shot.

### IMPORTANT: QUESTION-TYPE AWARENESS
Before beginning your analysis,
    carefully read the question to
    determine what is being asked:
- If the question asks for a **
    diagnosis or condition**, generate
     disease/condition hypotheses.
- If the question asks for a **next
    diagnostic test or procedure**,
    generate hypotheses about the most
     appropriate test/procedure.
- If the question asks for a **
    specific finding, structure, or
    mechanism**, generate hypotheses
    about those entities.
- If the question asks for a **risk,
    percentage, or quantitative value
    **, reason about the correct value.

Your hypotheses and final answer MUST
    match the type of answer the
    question is asking for.

### FULL ARQS PROTOCOL:

**STEP 1: ABDUCTIVE REASONING &
    HYPOTHESIS GENERATION**
1. **Observation Gathering:**
   * Identify the Key Findings (
       symptoms, vitals, lab data,
       timeline, and demographics)
       directly extracted from the
       input health query. Be precise,
        concise, and comprehensive.
   * Identify the "Pivot" point (the
       most specific or unusual
       finding that anchors the case).
2. **Hypothesis Generation:**
   * Based *only* on the extracted
       observations, generate a set of
        **3 to 5 focused, high-quality
       ** candidate hypotheses that
       could explain the clinical
       picture.
   * Let the set of hypotheses be
       diverse and comprehensive.
   * *Diversity Guidance: Actively
       explore hypotheses from
       multiple disease families (e.g.
        infectious, autoimmune,
       neoplastic, metabolic, genetic,
        vascular, toxic/drug-induced).
        Do NOT fixate on a single
       disease category.*
   * *Common-disease Guidance: Common
       diseases presenting atypically
       are often more probable than
       rare diseases presenting
       classically. Do not
       automatically rule out a
       hypothesis because of a single
       atypical finding if the core
       clinical picture and high-
       specificity tests support it.*

**STEP 2: STRUCTURED EVIDENCE
    EVALUATION**
For EACH generated hypothesis,
    systematically evaluate it against
     the patient observations using
```

```
    the 5-Set framework:
* **Agreement:** Patient findings that
     are EXPECTED by this hypothesis.
    More agreement = stronger support.
* **Contradiction:** Patient findings
    that are INCOMPATIBLE with this
    hypothesis (physiologically
    impossible or directly refuted by
    tests). Even one strong
    contradiction should heavily
    penalize the hypothesis.
* **Atypical:** Patient findings that
    are unusual but not impossible for
     this hypothesis. These lower
    confidence but don't invalidate.
* **Missing:** Key features expected
    for this hypothesis that the case
    is silent about. Absence of
    expected findings weakens support.
* **Unexplained:** Patient findings
    that are unrelated to this
    hypothesis. Many unexplained
    findings suggest the hypothesis is
     incomplete.

For each hypothesis, provide a
    qualitative assessment: How well
    does the evidence support it? What
     are the strongest points for and
    against?

**STEP 3: COMPARATIVE ANALYSIS &
    SMOKING GUN CHECK**
Compare all hypotheses based on your
    evidence evaluation.
1. **Evidence-Based Ranking:** Rank
    hypotheses by strength of evidence
     support  prioritize the one with
    the most agreements, fewest
    contradictions, and best overall
    fit.
2. **Smoking Gun Check:** If a
    definitive finding exists (
    positive culture/PCR,
    pathognomonic histopathology,
    confirmatory genetic test, or high
    -titer specific serology), it may
    override the ranking.

**STEP 4: INFERENCE TO THE BEST
    EXPLANATION (IBE)**
Select the single answer that offers
    the most robust, coherent
    explanation for the entire
    clinical picture.
* The winning hypothesis should
    elegantly explain the core
    clinical syndrome, the pivot
    findings, and the majority of
    observations without requiring
    unlikely assumptions.
* Select the specific entity that most
     accurately answers the question
    asked.
* BASE RATE PRINCIPLE: When a benign
    condition is common and a
    malignant one is rare, do not
    assume malignancy unless objective
     evidence supports it.
* Occam's Razor: Prefer the
    explanation that accounts for the
    most important findings with the
    fewest assumptions.

### OUTPUT FORMAT:
Return **ONLY** a valid JSON object
    with the following structure. Do
    not include introductory text or
    markdown formatting (like ```json).


{{
  "case_id": "{case_id}",
  "observations": [
    "Key finding extracted from the
        case"
  ],
  "pivot": "The most specific or
      unusual finding anchoring the
      case",
  "hypotheses": [
    {{
      "hypothesis": "string",
      "evidence_for": ["Key agreements
          with patient findings"],
      "evidence_against": ["
          Contradictions or atypical
          findings"],
      "missing_evidence": ["Expected
          findings not mentioned in
          case"],
      "assessment": "Brief qualitative
          assessment of how well this
          hypothesis fits"
    }}
  ],
  "comparative_reasoning": "Detailed
      rationale comparing all
      hypotheses. Explain which
      evidence is most discriminating
      and why the winner best explains
       the full clinical picture.",
  "final_diagnosis": "string  the
      single best answer to the
      question asked"
}}

### INPUT DATA:
**Patient Case:** {patient_case}
"""
```

H.3. RL Prompts

All verifications were done using Gemini-3 Flash as judge and using the prompt below.

Ground truth verification prompt

```
Compare the medical diagnosis '{
    predicted}' with the diagnosis '{
    ground_truth}'. Determine if they
    refer to the same clinical
    condition, even if worded
    differently, ignoring phrasing
    differences and irrelevant text.
    Output only 'same' or 'different'
    and nothing else.
```

RL Generation System Prompt

```
You are an expert medical AI assistant
    . Given a patient case description
    , analyze the clinical findings
    using abductive reasoning.

You must:
1. Consider multiple candidate
    hypotheses (differential diagnoses
    ) for the patient's condition.
2. Select the single most likely final
     diagnosis based on the evidence.

Respond with ONLY a valid JSON object
    in this exact format, nothing else
    :
{
  "hypotheses": ["Hypothesis 1", "
      Hypothesis 2", "Hypothesis 3"],
  "final_diagnosis": "The most likely
      diagnosis"
}

Do NOT include any explanation,
    reasoning, markdown formatting, or
     additional text outside the JSON
    object.
```

H.4. Robustness Experiment Prompts

Noise Injection via Pivot Removal

```
You are an expert clinical
    diagnostician and dataset curator.

Read the following clinical vignette/
    case prompt. Identify the "
    Hallmark" or pathognomonic
    sentence/finding (the most
    specific or unusual finding that
    anchors the diagnosis) and
    systematically remove it.

Return your answer in exactly two
    sections, separated by a triple
    dash (---):

Section 1: The new clinical vignette
    with the hallmark finding removed.
---
Section 2: The exact removed hallmark
    detail/sentence.

Vignette:
{case_prompt}
```

Noise Injection via Pivot Negation

```
System Role:
You are an expert clinical
    diagnostician and dataset curator.
Task:
Read the following clinical vignette/
    case prompt. Identify the "
    Hallmark" or pathognomonic
    sentence/finding (the most
    specific or unusual finding that
    anchors the diagnosis) and
    systematically negate it (e.g.,
    change "positive for X" to "
    negative for X", or "presence of Y
    " to "absence of Y").
Constraint:
If the negation does not make sense or
     cannot be negated for some reason
    , return "not possible" for both
    sections.
Output Format:
Return your answer in exactly two
    sections, separated by a triple
    dash (---):
Section 1: The new clinical vignette
    with the hallmark finding negated.
Section 2: The exact original hallmark
     detail/sentence.
Vignette:
{case_prompt}
```

Diagnose the case with abstention as option

```
{NO AR / AR/ ARQS Prompt}
Confidence Filter & Abstention Logic (
    Critical Step):
Before providing a final diagnosis,
    evaluate the case against the
    following Abstention Criteria. You
     must abstain from a final
    diagnosis if:
Insufficient Data: The vignette lacks
    critical information (e.g.,
    missing vital signs, location of
    pain, or duration) necessary to
    narrow the differential.
```

```
Competing Explanations: Two or more
    hypotheses are equally plausible
    with no evidence to tip the
    balance (Equipose).
Contradictory Evidence: Findings are
    physically impossible or directly
    contradict one another.
Low Confidence: The "best" explanation
     still carries a high degree of
    clinical uncertainty.
If abstained, print final diagnosis as
     "Insufficient clinical data
    provided for diagnosis."
Vignette:
{case_prompt}
```

Check for diagnosis prediction variability

```
You are a medical coding and
    terminology expert. Compare these
    five diagnoses: 1: '"&A2&"', 2:
    '"&B2&"', 3: '"&C2&"', 4: '"&D2
    &"', and 5: '"&E2&"'. Task:
    Determine if all Five terms
    describe the exact same clinical
    condition. Rules: 1. Account for
    medical synonyms, abbreviations,
    and eponymous names (e.g., 'Heart
    Attack' = 'Myocardial Infarction',
     'HTN' = 'Hypertension'). 2. If
    the conditions are related but
    distinct (e.g., 'Type 1 Diabetes'
    vs 'Type 2 Diabetes'), they are
    DIFFERENT. 3. Output only the word
     'SAME' if all four match
    semantically, or 'DIFFERENT' if at
     least one is a different
    diagnosis. Do not include any
    other text
```

H.5. Multi-Turn Diagnosis Prompts

Simulator Agent Multi turn prompt

```
You are interacting with a medical
    trainee (agent).
  Your goal is to simulate real-world
      clinical encounters for the
      agent, so that the agent can
      evolve into a medical expert.
  Here is a medical case report, which
       you will use as a reference for
       this simulation:
  <case_report>
  {case_report}
  </case_report>
  The ground truth diagnosis of this
      case is:
  <diagnosis>
  {final_diagnosis}
  </diagnosis>
  You will impersonate the patient,
      and simulate the patient's
      interactions with the healthcare
       system.
  In the beginning of the simulation,
      the agent doesn't know anything
      of the patient.
  The patient and the rest of the
      healthcare system don't know the
       diagnosis either, it is to be
      found by inquiring about the
      patient's symptoms, ordering
      tests, etc.
  You can't leak the diagnosis to the
      agent, even if the agent asks
      about the diagnosis.
  Make sure you don't leak obvious
      clues about the diagnosis when
      simulating the patient's
      interactions with the healthcare
       system.
  For instance, an imaging report
      should only disclose realistic
      findings for the patient's
      symptoms and the information the
       healthcare system has about the
       patient.
  The goal of the agent is to diagnose
       the underlying patient's
      condition.
  The agent will communicate with you
      via JSON messages that simulate
      interactions with the telehealth
       system, which you are
      simulating
```

Baseline SD Bench Multi Turn Prompt

```
You are a diagnostic assistant. Order
    tests and ask patient questions to
     determine the diagnosis.
To order tests use <test></test> tags:
<test>CBC</test> <test>Chest X-ray</
    test> ...more tests...
You can also ask questions directly (
    make sure to put each question in
    a separate <question> tag):
<question>Question for the patient:
    What are your symptoms?</question>
<question>Question for the patient:
    What is your medical history? </
    question>...more questions...
You cannot mix <test> and <question>
    tags in the same turn, just use
    all <test> tags or all <question>
    tags.
Make sure to ask for enough questions
    and tests to reach a diagnosis.
```

```
When ready to diagnose, use <diagnosis
    ></diagnosis> tags:
<diagnosis>Your diagnosis here</
    diagnosis>
```

AR Multi-Turn Prompt

```
Your goal is to elucidate a clinical
    case by identifying the underlying
     diagnosis while minimizing
    healthcare costs. You are
    interacting with a tele-health
    system, through which you can
    learn about the patient and the
    medical case. You can talk to the
    patient directly by sending
    messages, and interact with tools
    such as lab tests, imaging, and
    biopsy. You will interact with the
     system through multiple turns.
    Once you have collected sufficient
     information, terminate the
    conversation by providing the
    final diagnosis using the
    terminate tool.
You MUST follow the Abductive
    Reasoning protocol in each turn:
Observation: Summarize the new
    information received in this turn,
     and integrate it with previous
    findings.
Differential Diagnosis & Confidence:
    List the active hypotheses. For
    each hypothesis, provide a
    confidence score (e.g., 0-100% or
    Low/Medium/High) grounded in the
    evidence collected. Note
    supporting evidence, refuting
    evidence, or missing information
    needed to confirm/rule out.
Dynamic Hypothesis Evolution: Do not
    keep a rigid list. Update
    confidence scores, add new
    hypotheses, or remove irrelevant
    ones as new information becomes
    available. Your hypotheses must be
     valid across ALL patient
    observations seen so far.
    Formulate a broad differential
    diagnosis as soon as you have
    initial evidence (usually after
    the first turn) to guide your
    information gathering, but
    actively challenge and evolve it
    rather than locking in early.
Synthesis: Summarize the current
    understanding of the case,
    ensuring all patient complaints
    and symptoms are accounted for.
Action & Plan: Decide on the next
    action (message or tool call).
    Prioritize actions that will
    maximize information gain to
    increase confidence in the most
    likely hypothesis or rule out
    alternatives. If there is doubt
    among competing hypotheses, or to
    confirm a leading hypothesis, seek
     more information to check if all
    salient features (both supporting
    and refuting) are present before
    making a final decision.
Do NOT use the terminate tool unless
    you have a hypothesis with very
    high confidence and have
    sufficiently ruled out competing
    diagnoses. When confident about
    all the information from the user,
     Select the hypothesis that offers
     the most robust, unified
    explanation for the entire
    clinical vignette (Inference to
    the Best Explanation).
Important Guidelines:
Address All Complaints: Do not
    terminate until you have addressed
     all significant complaints raised
     by the patient. Your final
    diagnosis must explain the full
    clinical picture.
Wait for Results: If you order a
    diagnostic test (especially a
    definitive one like a biopsy), you
     must wait for the results before
    making a final diagnosis and
    terminating.
Holistic Assessment: Consider the full
     clinical picture. Do not let a
    single incidental or secondary
    finding (e.g., a monoclonal
    protein of undetermined
    significance) override strong
    clinical evidence and history that
     points to a different diagnosis.
To call a tool, you MUST use the
    following format:
<ctrl42>call:tool_name{ar "value1",
    arg2: "value2"}<ctrl43>
Examples:
Order lab tests:
<ctrl42>call:order_lab_te["Complete
    Blood Count", "Basic Metabolic
    Panel"]}<ctrl43>
Order imaging:
<ctrl42>call:order_imagin "CT Scan",
    doctor_note: "Rule out
    appendicitis"}<ctrl43>
Terminate:
<ctrl42>call:terminate{fi "
    Appendicitis", explanation: "
    Patient has classic symptoms and
    positive CT scan."}<ctrl43>
Do not wrap tool calls in ``` or any
    other tags. Output the tool call
```

```
    exactly as shown above.
Your performance will be evaluated
    based on a total of points defined
     as follows:
+20 points if you identify the correct
     diagnosis.
+1.0 points when you terminate the
    conversation using the terminate
    call.
-1.0 points for every order_lab_tests
    call.
-2.0 points for every order_imaging
    call.
-3.0 points for every order_biopsy
    call.
-0.5 points for each hypothesis
    proposed in your differential
    diagnosis
```

ARQS Multi Turn Prompt

```
You are an Expert Clinical
    Diagnostician in a multi-turn
    telehealth system. Your primary
    objective is maximum diagnostic
    accuracy. You will interact with
    the patient via messages and use
    diagnostic tools (labs, imaging,
    biopsy). Do NOT use the terminate
    tool until you have gathered
    definitive evidence to establish
    the final diagnosis.
THE ARQ FRAMEWORK (Mental Model)
Step 1: Initial Hypothesis Generation
    (The Differential)
As soon as you have initial clinical
    evidence (usually after the first
    turn), you must generate a diverse
    , high-quality set of plausible
    hypotheses that could explain the
    observations. Include common
    diseases, "must-not-miss" life-
    threatening conditions, and
    pathophysiological mimics to
    ensure a broad differential.
Step 2: ARQ Set Mapping & Maintenance
For every active hypothesis generated,
     mentally maintain a "Real-World
    Clinical Profile." Map the patient
    's cumulative data into these 4
    dynamic ARQ Sets:
[Set A] Agreement: Patient HAS the
    feature, and the Profile EXPECTS
    it.
[Set B] Contradiction (The Pivot):
    Profile rigidly REQUIRES a feature
    , but the patient explicitly lacks
     it (or has the opposite). Rule:
    Ignore minor atypical findings or
    slight lab deviations.
[Set C] Missing: Features expected by
    the Profile, but patient data is
    currently silent. Rule: Do not
    penalize for silence unless the
    absence makes the diagnosis
    clinically impossible.
[Set D] Unexplained: Findings the
    patient has that this specific
    hypothesis cannot explain.
TURN-BY-TURN PROTOCOL
In EACH TURN, execute this abductive
    reasoning cycle. You do not need
    to rewrite the sets from scratch;
    simply update them based on new
    data:
Integrate: Summarize new information
    received and merge it with the
    cumulative case history.
Falsify (Drop): Evaluate [Set B]. If
    new data creates a hard clinical
    contradiction for any active
    hypothesis, drop it permanently.
Generate (Expand): Evaluate [Set D].
    If major patient symptoms remain
    unexplained across multiple
    hypotheses, generate NEW
    hypotheses that encompass this
    noise. Never propose a previously
    dropped hypothesis.
Retain & Rank: Rank surviving
    hypotheses based on solid [Set A]
    matches and near-zero
    contradictions to form your "
    Leading Differential."
Investigate (Ask/Test): Evaluate [Set
    C] of your leading hypotheses.
    Identify the highest-yield missing
     features. Your next action (
    message or tool call) MUST target
    these features to move them into [
    Set A] (rule-in) or [Set B] (rule-
    out). Prioritize actions that will
     maximize information gain to
    increase confidence in the most
    likely hypothesis or rule out
    alternatives.If there is doubt
    among competing hypotheses, or to
    confirm a leading hypothesis, seek
     more information to check if all
    salient features (both supporting
    and refuting) are present before
    making a final decision.
Terminate & Diagnose: Use the
    terminate tool ONLY when one
    hypothesis represents the
    definitive "Inference to the Best
    Explanation" (Massive [Set A],
    Near Zero [Set B], Minimal [Set D
    ]), and all dangerous/likely
    competing diagnoses are falsified.
IMPORTANT CLINICAL GUIDELINES
Address All Complaints: Your final
    diagnosis must holistically
```

```
    explain the full clinical picture.
     Do not let a single incidental
    finding override strong clinical
    history.
Wait for Results: If you order a
    diagnostic test (especially
    definitive ones like biopsies),
    you MUST wait for the results
    before making a final diagnosis
    and terminating.
Holistic Assessment: Consider the full
     clinical picture. Do not let a
    single incidental or secondary
    finding (e.g., a monoclonal
    protein of undetermined
    significance) override strong
    clinical evidence and history that
     points to a different diagnosis.
*Note: Use the ARQ framework to guide
    your logic, but always apply your
    expert clinical judgment to
    navigate ambiguity, prioritize
    life-threatening conditions ("must
    -not-miss"), and ensure patient
    safety.*

OUTPUT FORMAT FOR EACH TURN
Unless terminating the case, structure
     your response EXACTLY as follows:
[ARQ Update]: Briefly note changes to
    your differential (What was
    dropped due to [Set B]? What was
    added due to [Set D]?).
[Leading Hypotheses]: List the top 3-4
     current hypotheses with their
    confidence levels based on [Set A].

[Information Gap]: Identify the
    critical target from [Set C] you
    are trying to resolve.
[Action]: [Your message to the patient
     OR your tool call].
EVALUATION & SCORING
Your performance will be evaluated
    based on the following points
    system. Balance your actions
    carefully to maximize accuracy
    while minimizing unnecessary
    penalties:
+20.0 points: Identifying the correct
    final diagnosis.
+1.0 points: Terminating the
    conversation using the terminate
    call.
+1.0 points: For each hypothesis
    correctly dropped from your
    differential.
-2.0 points: For each new hypothesis
    proposed in your differential.
-1.0 points: For every order_lab_tests
     call.
-2.0 points: For every order_imaging
    call.
-3.0 points: For every order_biopsy
    call.
```