

Supplementary Material

Table of Contents

A	Extended Related Work	8
B	Algorithm and Implementation Details	9
B.1	CURIOSITREE Algorithm	9
B.2	Implementing CURIOSITREE via Prompting	9
C	Clinical Diagnosis Simulator: Additional Details	10
D	Prompts Used in Experiments	11
D.1	Agent Prompts	11
D.1.1	General Prompts	11
D.1.2	Action Sampling Prompts	13
D.1.3	CURIOSITREE Prompts	15
D.1.4	Self-Evaluation Prompts	18
D.2	Environment Prompts	19
E	Additional Experiments and Results	20
E.1	Correspondence with Human Intuition	20
E.2	Distribution of Action Types when Diagnosing “Lupus”	21
E.3	Sample History of Interactions from CURIOSITREE to Diagnose “Lupus”	21
E.4	Sample Dialogue from Institutional Knowledge	22
E.5	Results for Additional Diagnoses	22
E.5.1	Results for “Primary Biliary Cirrhosis”	23
E.5.2	Results for “Hypertension”	25
E.5.3	Results for “Type 2 Diabetes Mellitus”	27
E.5.4	Results for “Asthma”	29
E.5.5	Results for “Rheumatoid Arthritis”	31
E.5.6	Results for “Multiple Sclerosis”	33
E.5.7	Results for “Cystic Fibrosis”	35
E.5.8	Results for “Amyotrophic Lateral Sclerosis”	37
E.5.9	Results for “Myasthenia Gravis”	39

A. Extended Related Work

Active Learning and Optimal Experiment Design. Like active learning (Cohn *et al.*, 1996; Gal *et al.*, 2017; Y. Xia *et al.*, 2025), we study the problem of optimally acquiring information to improve predictive accuracy; however, unlike the classical setting, which permits downstream model training and assumes a moderate-to-large dataset, we operate at test-time, and study *instance-wise acquisition*, choosing to gather information about each datum rather than learning patterns between data points. In this way, elements of our approach can be viewed through the lens of optimal experimental design (Chaloner and Verdinelli, 1995; Rainforth *et al.*, 2024; Piriyaikulij *et al.*, 2024), simulating experiments and outcomes and choosing to perform those with the greatest informative value.

Tool Use in LLMs. A substantial body of work implements mechanisms of querying individual partitions of the Information Ecosystem to augment the information available to large language models at test time. Methods for generating deliberative reasoning tokens (Wei *et al.*, 2022; Yao *et al.*, 2023), querying documentary knowledge (P. Lewis *et al.*, 2020; Asai *et al.*, 2023; Thulke *et al.*, 2024), asking informative questions (below), and designing experimental parameters (Ghafariollahi and Buehler, 2024; R. Li *et al.*, 2024; Roohani *et al.*, 2024). Despite this, there has been comparatively little study of how to optimally balance between query mechanisms—especially those of varying cost. Although much of the foundational work in LLM tool use is built on in-context learning (Schick *et al.*, 2023; Hu *et al.*, 2023; Gou *et al.*, 2023), in-context learning is often insufficient (Huang *et al.*, 2023), and that incorporating simulation- and heuristic-based approaches may offer a solution (B. Wang *et al.*, 2024).

Question-Asking in LLMs. Several recent works have studied the task of asking informative questions using language models. Standard training-time methods have included fine-tuning with preference optimization (Mazzaccara *et al.*, 2024; M. J. Zhang *et al.*, 2024; Lee *et al.*, 2025) or self-play (Andukuri *et al.*, 2024); some recent work has sought to benchmark the performance of question-asking LLMs in settings of insufficient information (W. Wang *et al.*, 2024; B. Z. Li *et al.*, 2025; S. S. Li *et al.*, 2025). One prominent test-time approach is Uncertainty-of-Thoughts, which attempts to ask the most informative binary question by computing a closed-form expression for the expected information gain over a number of candidate options. Although elements of our work conceptually draw inspiration from Uncertainty-of-Thoughts, our contribution substantially extends their framework by providing a means of accommodating a diverse set of possible actions with open-ended consequences (beyond the “yes/no” questions that are the primary focus of their study), performing cost-sensitive action selection, and by leveraging self-assessed confidence scores to relax their assumption of a uniform distribution over predictions at each point in time.

B. Algorithm and Implementation Details

B.1. CURIOSITREE Algorithm

Algorithm 1 CURIOSITREE: Strategic, Test-Time Information Acquisition

Require: Initial covariates $X_{d_0}^{(i)}$, predictive functions $\mathcal{G} = \{g_{d_t}\}$, environment simulator $\widehat{\mathcal{E}}$, cost function c , prediction threshold τ , budget B , utility-cost tradeoff λ , top- k size k , number of action samples k' .

- 1: $C \leftarrow 0$
- 2: $t \leftarrow 0$
- 3: **while** true **do**
- 4: Compute predictive distribution $g_{d_t}(X_{d_t}^{(i)})$ with top- k predictions $\{\widehat{y}_{t,j}^{(i)}, \widehat{p}_{t,j}^{(i)}\}_{j=0}^{k-1}$
- 5: **if** $\max_j \widehat{p}_{t,j}^{(i)} \geq \tau$ **then**
- 6: **return** $\widehat{Y}^{(i)} \leftarrow \arg \max_j \widehat{p}_{t,j}^{(i)}$
- 7: **end if**
- 8: Sample k' actions $\{\tilde{a}_t^{(j)}\}_{j=0}^{k'-1} \subseteq \mathcal{A}_t$ that do not exceed the budget: $c(\tilde{a}_t^{(j)}) + C \leq B$
- 9: **for** each candidate action $\tilde{a}_t^{(j)}$ **do**
- 10: **for** each top- k prediction $\widehat{y}_{t,l}^{(i)}$ **do**
- 11: Simulate new covariates $\tilde{X}_{d_{t+1}}^{(i)} \sim \widehat{\mathcal{E}}(X_{d_t}^{(i)}, \tilde{a}_t^{(j)}, \widehat{y}_{t,l}^{(i)})$
- 12: Compute updated prediction $g_{d_{t+1}}(\tilde{X}_{d_{t+1}}^{(i)})$
- 13: Evaluate logical consistency mask $A_{t,l}^{(j)}$ over top- k predictions
- 14: **end for**
- 15: Compute surrogate EIG as,
- 16:
$$\widehat{\text{EIG}}(\tilde{a}_t^{(j)} \mid X_{d_t}^{(i)}) = \begin{cases} -\log \left(\frac{1}{k} \sum_{l=0}^{k-1} A_{t,l}^{(j)} \right) & \text{if } \sum_{l=0}^{k-1} A_{t,l}^{(j)} > 0 \\ 0 & \text{otherwise.} \end{cases}$$
- 17: **end for**
- 18: Select action: $a_t \leftarrow \arg \max_{\tilde{a}_t^{(j)}} \left[\widehat{\text{EIG}}(\tilde{a}_t^{(j)} \mid X_{d_t}^{(i)}) - \lambda c(\tilde{a}_t^{(j)}) \right]$
- 19: Update covariates: $X_{d_{t+1}}^{(i)} \leftarrow \mathcal{E}(X_{d_t}^{(i)}, a_t)$
- 20: $C \leftarrow C + c(a_t)$
- 21: $t \leftarrow t + 1$
- 22: **end while**
- 23: **return** $\widehat{Y}^{(i)} \leftarrow \emptyset$ ▷ Abstain if budget exhausted.

B.2. Implementing CURIOSITREE via Prompting

Implementing $\widehat{\mathcal{E}}$ via Prompting. Recall that $\widehat{\mathcal{E}}$ represents the environment simulator for the purpose of evaluating Equation 4. In theory, the implementing language model should emit responses roughly proportionally consistent with the present set of hypotheses. In practice, the language model emissions are of considerably lower-entropy, which worsens the quality of the approximation. To we circumvent this challenge with a strategy we call *prior locking*, where we explicitly compute the expectation in Equation 3 as

$$\frac{1}{k} \sum_{j=1}^k \mathbb{E}_{\tilde{X}_{d_{t+1}}^{(i)} \sim \widehat{\mathcal{E}}(X_{d_t}^{(i)}, a_t, \widehat{y}_{t,j}^{(i)})} \left[H \left(\widehat{\text{Pr}}(Y^{(i)} \mid \tilde{X}_{d_{t+1}}^{(i)}) \right) \right], \quad (5)$$

and explicitly condition each call to the environment simulator by a candidate label $\widehat{y}_{t,j}^{(i)}$, representing one of the top- k model predictions at time t .

Implementing \mathcal{G} via Prompting. By implementing each $g_{d_t} \in \mathcal{G}$ as a generative process using top- k sampling and constrained decoding (Deutsch *et al.*, 2019; Kuchnik *et al.*, 2023), our method does not assume access to an exhaustive label set. Our prompt for each g_{d_t} accepts as input the observed history of covariates and actions, \mathcal{H}_t , and requests a structured list of (a) the k most probable values of the distribution given the current history, $\widehat{y}_{t,0}^{(i)}, \dots, \widehat{y}_{t,k-1}^{(i)}$, and (b) the probability

scores corresponding to each value, $\tilde{p}_{t,0}^{(i)}, \dots, \tilde{p}_{t,k-1}^{(i)}$. Although constrained decoding guarantees that exactly k items and scores are generated, and that the scores are nonnegative floating point numbers, it does not guarantee that the distribution is normalized. We therefore normalize the distribution by dividing each score by the sum of scores assigned across the top- k classes, $\hat{p}_{t,j}^{(i)} = \tilde{p}_{t,j}^{(i)} / \sum_{l=0}^{k-1} \tilde{p}_{t,l}^{(i)}$.

This setup makes it challenging to directly evaluate Equation 3, because (a) these top- k distributions are necessarily sparse and unnormalized, meaning we have no closed-form distribution over which we can compute entropy, and (b) because we want Equation 3 to reflect the *semantic distribution of candidate labels* in accommodating cases where probability mass is distributed over distinct-but-semantically-similar concepts. We therefore use a surrogate of Equation 3 that rewards actions based on the proportion of present candidates that they eliminate from consideration. Using constrained decoding, our prompt to implement this accepts as input the current top- k list of candidates $Y^{(i)}$, the action a_t , and the sampled response $\tilde{X}_{d_t+1}^{(i)}$, and produces a structured boolean array $A_t^{(i)} \in [0, 1]^k$ of k booleans indicating whether each element in the list of candidates remains logically consistent given the action and the response (1 if `True`). Then, we implement Equation 4 using $\widehat{\text{EIG}}(a_t | X_{d_t}^{(i)}) = -\log \left[(1/k) \sum_{l=0}^{k-1} A_{t,l}^{(i)} \right]$, with $\widehat{\text{EIG}}(a_t | X_{d_t}^{(i)})$ set to 0 if no options are consistent. In a sense, this approach implicitly assumes that the true label is among the top- k candidates and rewards actions that are likely to disambiguate among them. Although this approximation introduces potential bias when the ground truth label is not represented among the current predictions, empirically, we find it yields an effective search heuristic.

C. Clinical Diagnosis Simulator: Additional Details

Our clinical diagnosis simulator provides the following four levers of information acquisition. For the purposes of this testbed, the cost assigned to each class of action is largely arbitrary, though their relative cost-ordering approximately reflects realistic clinical practice. In reality, the exact cost values would be determined by a combination of financial cost (*e.g.*, token generation may be inexpensive; running a laboratory test may be expensive) and other considerations (*e.g.*, limited availability for a certain form of laboratory test).

1. **Generate Reasoning Tokens** (Cost: 1). The clinician can generate intermediary reasoning tokens regarding the patient’s condition. These tokens are added to the history so that future predictions and actions are conditional on these tokens. Because reasoning is an intrinsic process, no corresponding response is provided by the experiment simulator.
2. **Perform RAG on Wikipedia** (Cost: 1). The clinician can generate a sequence of tokens corresponding to a query for information. In response to the information query, the experiment simulator matches the query to the p most similar chunks of text from Wikipedia using a search over the `NeuML/txtai-wikipedia` embeddings. The text of the chunks is added to the history.
3. **Ask the Patient a Question** (Cost: 2). The clinician can generate a sequence of tokens representing a question for the patient. In response, the experiment simulator instantiates a language model that has been prompted to respond as a patient whose ground-truth diagnosis is $Y^{(i)}$. This language model is prompted to respond with ignorance if the question asked by the clinician is too technical or specific. The question and response are then added to the history. This baseline is most conceptually similar to an open-ended version of Uncertainty-of-Thoughts.
4. **Requisition a Laboratory Test** (Cost: 3). The clinician can generate a sequence of tokens representing a specific laboratory assessment to be performed on the patient. In response, the experiment simulator instantiates a language model that has been prompted to yield the results of a laboratory test corresponding to a patient with a ground-truth diagnosis is $Y^{(i)}$. The request and test result are then added to the history.

No initial information is provided to the clinician at the start of simulation. For each of the four classes of action, we generate five action candidates for a total of twenty actions evaluated each turn. Each action is scored, either using the CURIOSITREE heuristic (Equation 4) or one of the baseline heuristics described below, and the highest-scored action is passed to the environment for evaluation. The simulation is terminated under only two conditions. First, if the cumulative cost incurred exceeds the budget, B : we track this as an abstention ($\hat{Y}^{(i)} = \emptyset$), because insufficient information was gathered to make a confident prediction. Second, if a prediction for $\hat{Y}^{(i)}$ is rendered according to Algorithm 1, Line 6: if the prediction is correct (adjudicated using LLM-as-a-judge (Zheng *et al.*, 2023)), we record this as a successful diagnosis, and if the prediction is incorrect, we record this as an unsuccessful diagnosis.

D. Prompts Used in Experiments

D.1. Agent Prompts

D.1.1. GENERAL PROMPTS

Context Prompt

Description: *The context prompt is prepended to the prompts wherein (a) the agent creates the predictive distribution or, (b) samples actions. The purpose of this prompt is to provide the agent with context about the current state. This prompt is generated dynamically based on the current. The below pseudocode illustrates how the prompt is generated at each turn.*

```
def generate_prompt(history):
    # history is a dictionary with keys:
    #   "actions" (list of strings)
    #   "outcomes" (list of strings)
    #   "action_types" (list of strings), each element is one of: "question",
    #   ↪ "experiment", "RAG", "reasoning"

    if history["actions"] is empty:
        # No previous actions taken
        return [
            {
                "role": "system",
                "content": "You are a clinician seeing a patient. You are attempting
                ↪ to elicit their primary diagnosis (specific disease name). You
                ↪ do not yet know any information about the patient."
            }
        ]
    else:
        # Start with system message summarizing current clinical context
        context = [
            {
                "role": "system",
                "content": "You are a clinician seeing a patient. You are attempting
                ↪ to elicit their primary diagnosis (specific disease name). You
                ↪ have currently undertaken the following assessments, and
                ↪ obtained the following outcomes."
            }
        ]

        # Iterate over past interactions
        for i from 0 to length(history["actions"]) - 1:
            action = history["actions"][i]
            outcome = history["outcomes"][i]
            action_type = history["action_types"][i]

            if action_type == "question":
                context.append({
                    "role": "assistant",
                    "content": "You asked the patient, " + action + "? \n\nThe
                    ↪ patient responded, " + outcome + "."
                })
            else if action_type == "experiment":
                context.append({
                    "role": "assistant",
                    "content": "You ordered the test, " + action + ". \n\nThe result
                    ↪ indicated, " + outcome + "."
                })
            else if action_type == "RAG":
                context.append({
```

```

        "role": "assistant",
        "content": "You performed a Wikipedia search, consisting of the
        ↳ following search query and document retrieval. \n\n" +
        ↳ action
    })
    else if action_type == "reasoning":
        context.append({
            "role": "assistant",
            "content": "You reasoned to yourself that, '" + action + "'."
        })

    return context

```

Prediction Prompt

Description: This prompt requires the agent to produce a list of the top- k most likely diagnoses given its current knowledge (from the Context Prompt). This prompt is used in conjunction with structured decoding to guarantee the output format.

```

def generate_prediction_prompt(k):
    return [
        {
            "role": "system",
            "content": "Based on your current knowledge, provide " + str(k) + " of
            ↳ your best guesses of the patient's diagnosis. Format your output as
            ↳ a list of tuples, (x_i, s_i), where x_i is a string representing a
            ↳ guess, and s_i is a number BETWEEN 0 AND 1 representing the
            ↳ probability that x_i is the patient's true diagnosis. YOU MUST
            ↳ PROVIDE YOUR BEST GUESSES; THESE GUESSES MUST BE UNIQUE DIAGNOSES.
            ↳ EVEN IF YOU DO NOT HAVE ENOUGH INFO, PROVIDE GUESSES. YOU MUST
            ↳ PROVIDE GUESSES UNDER ALL CIRCUMSTANCES. Provide NO OTHER OUTPUT
            ↳ WHATSOEVER."
        }
    ]

```

D.1.2. ACTION SAMPLING PROMPTS

Generate Reasoning Action

Description: This prompt instructs the agent to generate internal reasoning statements to aid in diagnosis. The number of actions, k' , is a tunable parameter.

```
def generate_reasoning_prompt(k_prime):
    return [
        {
            "role": "system",
            "content": "Generate " + str(k_prime) + " independent logical reasoning
            ↳ statements that may help determine the patient's diagnosis. NO
            ↳ RESPONSE WILL BE PROVIDED TO THESE; THEY ARE FOR YOUR CONSIDERATION
            ↳ AND CONTEMPLATION ONLY. They are NOT to be questions. START EACH
            ↳ CHAIN WITH 'We know that,' and write one sentence summarizing what
            ↳ we know of the patient; then provide logical deduction from there.
            ↳ Format your output as a comma-separated list, SURROUNDING EACH
            ↳ REASONING CHAIN WITH PARENTHESES. Each reasoning chain should be a
            ↳ complete and coherent thought process. Provide NO OTHER OUTPUT."
        }
    ]
```

Generate Wikipedia RAG Search Queries

Description: This prompt instructs the agent to generate information retrieval queries to aid in diagnosis. The number of queries, k' , is a tunable parameter.

```
def generate_multiple_queries_prompt(k_prime):
    return [
        {
            "role": "system",
            "content": "You are permitted to generate information retrieval queries
            ↳ to help a clinician determine a patient's diagnosis. Ponder " +
            ↳ str(k_prime) + " Wikipedia search queries that could retrieve
            ↳ information relevant to diagnosing this patient. Format your output
            ↳ as a comma-separated list, surrounding each query with parentheses.
            ↳ EACH OUTPUT MUST BE A SINGLE QUERY. Provide NO OTHER OUTPUT."
        }
    ]
```

Generate Patient Questions

Description: This prompt instructs the agent to generate a set of diagnostic questions to ask the patient. The number of questions, k' , is a tunable parameter.

```
def generate_question_prompt(k_prime):  
    return [  
        {  
            "role": "system",  
            "content": "You are permitted to ask the patient a question -- the  
                ↳ patient will respond to the question truthfully to the best of their  
                ↳ knowledge. Ponder " + str(k_prime) + " different independent  
                ↳ questions to ask. ENSURE THAT QUESTIONS ARE A MIX OF YES/NO,  
                ↳ MULTIPLE-CHOICE, AND OPEN-ENDED QUESTIONS. Format your output as a  
                ↳ comma-separated list, surrounding each action with parentheses. EACH  
                ↳ OUTPUT MUST BE A SINGLE QUESTION. Provide NO OTHER OUTPUT."  
        }  
    ]
```

Generate Laboratory Test Requisitions

Description: This prompt instructs the agent to propose clinical tests or assessments. Each test should be a complete and specific requisition. The number of requisitions, k' , is a tunable parameter.

```
def generate_experiment_prompt(k_prime):  
    return [  
        {  
            "role": "system",  
            "content": "You can run a laboratory test on the patient. Write a  
                ↳ DETAILED REQUISITION for an assessment or test that you would run on  
                ↳ the patient. Consider " + str(k_prime) + " different tests to run.  
                ↳ Format your output as a comma-separated list, surrounding each  
                ↳ action with parentheses. EACH OUTPUT MUST BE A COMPLETE TEST  
                ↳ REQUISITION. Provide NO OTHER OUTPUT."  
        }  
    ]
```

D.1.3. CURIOSITREE PROMPTS

The first two prompts below present the prompts used in CURIOSITREE’s environment simulator, $\hat{\mathcal{E}}$. Because neither reasoning nor RAG requires interaction with the environment, these actions do not need to be simulated; rather, their EIG can be directly computed by adding the action to the history and treating the response as `null`.

Environment Simulator ($\hat{\mathcal{E}}$) – Patient Response to Question

Description: This prompt simulates a patient’s response to a clinician’s question, assuming the patient has a known underlying disease. Both the question and underlying disease are tunable parameters, which is useful for implementing this prompt in the context of prior locking.

```
def simulate_answers(question, underlying_diagnosis):
    return [
        {
            "role": "system",
            "content": "You are an oracle providing information to clinicians about
            ↪ their patients. The current patient has the disease " +
            ↪ underlying_diagnosis + ". If the 'user' (clinician) asks a question,
            ↪ you are to respond 'The patient responds,' and then respond to the
            ↪ question as a patient would, recognizing that the patient has no
            ↪ knowledge of specialized clinical terminology, ICD-10 codes, etc. In
            ↪ general, be as succinct and direct as possible."
        },
        {
            "role": "user",
            "content": question
        }
    ]
```

Environment Simulator ($\hat{\mathcal{E}}$) – Result of Laboratory Test

Description: This prompt simulates the result of a lab test based on a known diagnosis. The oracle returns a test result consistent with the underlying condition. Both the test description and underlying disease are tunable parameters, which is useful for implementing this prompt in the context of prior locking.

```
def simulate_exp_outcome(test_description, sample_prior):
    return [
        {
            "role": "system",
            "content": "You are an oracle providing information to clinicians about
            ↪ their patients. The current patient has the disease " + sample_prior
            ↪ + ". The 'user' (clinician) will order a lab test; you are to
            ↪ respond 'The result indicates,' and then provide a laboratory test
            ↪ result consistent with the patient's underlying diagnosis. In
            ↪ general, be as succinct and direct as possible."
        },
        {
            "role": "user",
            "content": test_description
        }
    ]
```

Generate Assignment Prompt

Description: This prompt determines which candidate diagnoses remain logically consistent with a given action and response. This is used for computing the surrogate expected information gain. The output is a list of booleans for each candidate diagnosis; this prompt is used in conjunction with structured decoding to enforce that a list of k booleans be generated as output.

```
def assignment_prompt(action_type, action, response, candidates):
    # action_type: one of ["question", "reasoning", "RAG", "experiment"]
    # action: string or tuple, depending on type; if the action_type is RAG, then
    #   → action is a 3-tuple consisting of the query, the retrieved document
    #   → identifier, and the retrieved document text.
    # response: string
    # candidates: list of candidate diagnosis strings

    assert action_type in ["question", "reasoning", "RAG", "experiment"]

    if action_type == "question":
        return [
            {
                "role": "system",
                "content": "Given the question, '" + action + "', and the response,
                → '" + response + "', determine which of the following diagnoses
                → remain LOGICALLY CONSISTENT with the response; a response that
                → is not relevant to a certain diagnosis does NOT NECESSARILY mean
                → that the diagnosis is incompatible with the retrieval. Format
                → your output as a comma-separated list of booleans, surrounding
                → each boolean with parentheses, e.g., (b_1), (b_2), .... If b_i is
                → TRUE, this indicates that candidate i MAY BE LOGICALLY
                → CONSISTENT with the response to the question. Produce NO OUTPUT
                → OTHER THAN THE LIST OF BOOLEANS. The candidates are: [" + ',
                → '.join(candidates) + "]."
```

```

        "content": "Given the information retrieval request, '" + action[0]
        ↪ + "'", and retrieved information '" + str((action[1], action[2]))
        ↪ + "'", determine which of the following diagnoses remain
        ↪ logically consistent with the retrieved information. Format your
        ↪ output as a comma-separated list of booleans, surrounding each
        ↪ boolean with parentheses, e.g., (b_1), (b_2), .... If b_i is TRUE,
        ↪ this indicates that candidate i MAY BE LOGICALLY CONSISTENT with
        ↪ the retrieved information. Be selective; ONLY MARK candidate i
        ↪ as False if the retrieved information CONTAINS RATIONALE THAT
        ↪ RENDERS IT INCONSISTENT; if the retrieved information is merely
        ↪ off-topic you must mark it as True. Produce NO OUTPUT OTHER THAN
        ↪ THE LIST OF BOOLEANS. The candidates are: [" + ',
        ↪ '.join(candidates) + "]."
    }
]

else if action_type == "experiment":
    return [
        {
            "role": "system",
            "content": "Given the experiment, '" + action + "'", and the outcome,
            ↪ '" + response + "'", determine which of the following diagnoses
            ↪ remains LOGICALLY CONSISTENT with the outcome of the experiment;
            ↪ an experiment that is not relevant to a certain diagnosis does
            ↪ NOT NECESSARILY mean that the diagnosis is incompatible with the
            ↪ experiment. Format your output as a comma-separated list of
            ↪ booleans, surrounding each boolean with parentheses, e.g.,
            ↪ (b_1), (b_2), .... If b_i is TRUE, this indicates that candidate
            ↪ i MAY BE LOGICALLY CONSISTENT with the outcome of the experiment.
            ↪ Produce NO OUTPUT OTHER THAN THE LIST OF BOOLEANS. The
            ↪ candidates are: [" + ', '.join(candidates) + "]."
        }
    ]

```


D.1.4. SELF-EVALUATION PROMPTS

Self-Evaluation Prompt

Description: This prompt asks the agent to evaluate multiple candidate actions based on their expected informativeness relative to their cost. Each action is assigned a score between 0 and 1.

```
def self_eval(actions, action_types, costs):
    # actions: list of action strings
    # action_types: list of corresponding types: "question", "experiment", "RAG",
    # ↪ "reasoning"
    # costs: dictionary mapping action types directly to their numerical costs

    actions_in_words = []
    for i from 0 to length(actions) - 1:
        a = actions[i]
        a_type = action_types[i]
        cost = costs[a_type]

        if a_type == "question":
            actions_in_words.append("(" + str(i) + "): Ask the patient, '" + a + "'
            ↪ (Cost: " + str(cost) + ")")
        else if a_type == "experiment":
            actions_in_words.append("(" + str(i) + "): Order the laboratory test, '"
            ↪ + a + "' (Cost: " + str(cost) + ")")
        else if a_type == "RAG":
            actions_in_words.append("(" + str(i) + "): Perform a Wikipedia search,
            ↪ consisting of the following query and document retrieval. \n\n" + a
            ↪ + " (Cost: " + str(cost) + ")")
        else if a_type == "reasoning":
            actions_in_words.append("(" + str(i) + "): Reason to yourself that, '" +
            ↪ a + "' (Cost: " + str(cost) + ")")

    all_actions_text = join(actions_in_words, "\n")

    return [
        {
            "role": "system",
            "content": "You have the opportunity to undertake ONE of the following "
            ↪ + str(length(actions)) + " assessments to determine the patient's
            ↪ diagnosis. Assign each a numeric score BETWEEN 0 AND 1 where a high
            ↪ score corresponds to a better (more informative) action in the
            ↪ current context, relative to its cost.\n\n" + all_actions_text +
            ↪ "\n\n Format your output as a comma-separated list of scores, (s_1),
            ↪ ..., (s_" + str(length(actions)) + "), surrounding each score with
            ↪ parentheses. s_i is the score assigned to the ith action. Provide NO
            ↪ OTHER OUTPUT."
        }
    ]
```

D.2. Environment Prompts

Oracle Response Prompt

Description: Generates the prompt for an oracle model simulating patient responses, lab test outcomes, or verifying the correctness of a prediction.

```
def generate_oracle_prompt(question, ground_truth, action_type):
    assert action_type in ["question", "experiment", "pred"]

    if action_type == "question":
        prompt = [
            {
                "role": "system",
                "content": "You are an oracle providing information to clinicians
↪ about their patients. The current patient has the disease " +
↪ ground_truth + ". The clinician will ask questions to ascertain
↪ the diagnosis; DO NOT EXPLICITLY MENTION '" + ground_truth + "'
↪ IN ANY ANSWER. Provide truthful answers with as little detail
↪ and as tersely as possible. \n\n Respond to the clinician with
↪ 'The patient responds,' followed by a response as a patient
↪ would, who lacks clinical terminology knowledge and may have
↪ some but not all symptoms of the disease. \n\n If the question
↪ is unrelated to the patient or diagnosis, respond with 'I don't
↪ know'."
            },
            {
                "role": "user",
                "content": question
            }
        ]

    else if action_type == "experiment":
        prompt = [
            {
                "role": "system",
                "content": "You are an oracle providing information to clinicians
↪ about their patients. The current patient has the disease " +
↪ ground_truth + ". The clinician will order lab tests to
↪ ascertain diagnosis; DO NOT EXPLICITLY MENTION '" + ground_truth
↪ + "' IN ANY ANSWER.\n\n For a test requisition given by the
↪ clinician, respond with 'The test yields,' followed by plausible
↪ test results consistent with the disease (not necessarily
↪ canonical). Do not interpret results; only provide specific
↪ biomarker/test values succinctly. \n\n If the request is
↪ unrelated to the patient or diagnosis, respond with 'I don't
↪ know'."
            },
            {
                "role": "user",
                "content": question
            }
        ]

    else if action_type == "pred":
        prompt = [
            {
                "role": "system",
```

```

        "content": "You are an oracle grading predictions with ground truth
        ↪ [" + ground_truth + "]. The user will ask: \"Is it 'X'?\" If 'X'
        ↪ exactly matches " + ground_truth + ", respond '[END -- success]'.
        ↪ Otherwise, respond '[END -- failure]'. NEVER provide any other
        ↪ output."
    },
    {
        "role": "user",
        "content": question
    }
]

return prompt

```

E. Additional Experiments and Results

E.1. Correspondence with Human Intuition

We present a simple validation of our heuristic by confirming that the utility assigned to four different questions of varying quality correlate with human intuition. We evaluate questions that are irrelevant (“Q1: What is the best brand of BBQ for an outdoor family event?”), overly-specific (“Q2: Is your diagnosis primary biliary cholangitis?”), weakly-informative (“Q3: Is the problem associated with your legs?”), and strongly-informative (“Q4: What brings you into the clinic today?”), and we compute EIG as described in Equation 3. We anticipate that the irrelevant and overly-specific questions are assigned low utility scores, while the weakly- and strongly-informative questions enjoy higher utility scores. The result in Figure 3 confirms that the EIG associated with each question corresponds largely with human intuition.

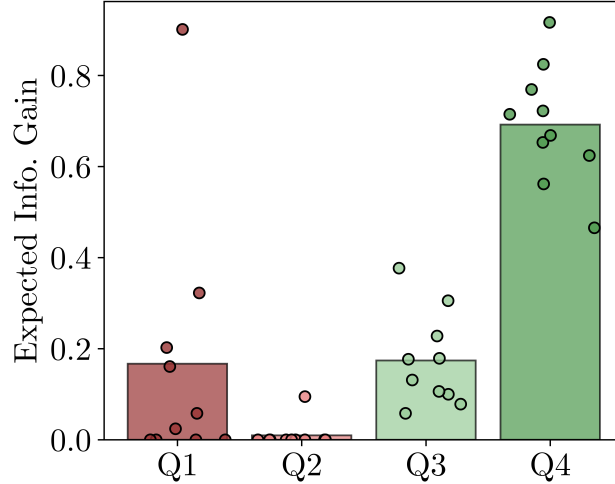


Figure 3: The orderings of Q1 through Q4 (Section E.1) using the CURIOSITREE utility function correspond with human intuition about the value of each question in the clinical diagnostic setting.

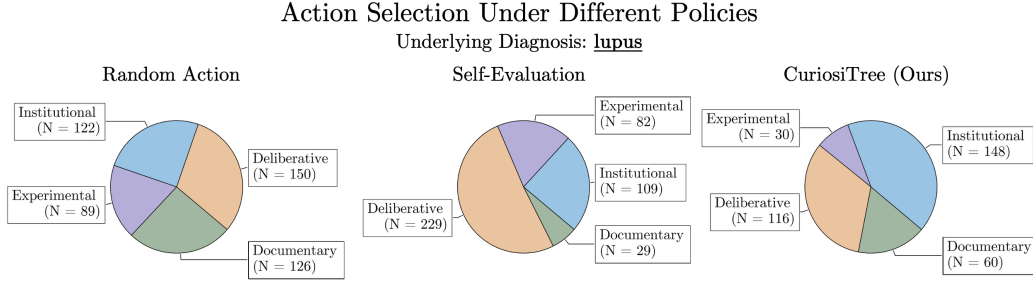


Figure 4: The action classes selected under different policies when diagnosing “lupus”. Random Action selection induces a more uniform distribution over action classes, while Self-Evaluation is dominated by Deliberative actions. CURIOSITREE balances between Institutional and Deliberative actions relatively evenly.

E.2. Distribution of Action Types when Diagnosing “Lupus”

E.3. Sample History of Interactions from CURIOSITREE to Diagnose “Lupus”

Successfully Diagnosing “Lupus” with CURIOSITREE

Description: This is a particularly interesting example in that CURIOSITREE receives limited information, and relies predominantly on subsequent reasoning steps to bootstrap the agent’s confidence level to the point where it can successfully render a diagnosis.

Clinician Asks: *What brings you to the clinic today?*

Response from \mathcal{E} : *The patient responds, I’ve been feeling really tired lately and my joints have been hurting a lot, especially in the morning.*

Clinician Asks: *Have you noticed any rashes or skin lesions on your body?*

Response from \mathcal{E} : *The patient responds, “Sometimes I get this weird butterfly-shaped rash on my face, especially after being in the sun.”*

Clinician Reasons: *We know that the patient’s symptoms are triggered by sun exposure. This is a common trigger for autoimmune diseases such as lupus or dermatomyositis.*

Response from \mathcal{E} : *N/A*

Clinician Reasons: *We know that the patient’s symptoms are triggered by sun exposure and include joint pain and a rash*

Response from \mathcal{E} : *N/A*

Clinician Reasons: *We know that the patient’s symptoms are triggered by sun exposure and include joint pain. This suggests a condition such as lupus or dermatomyositis that is exacerbated by UV light.*

Response from \mathcal{E} : *N/A*

Clinician Queries Wikipedia: *Search Query: Butterfly-shaped rash on face causes*

Response from \mathcal{E} : *Retrieval: [Malar rash] A malar rash, also called butterfly rash, is a medical sign consisting of a characteristic form of facial rash. It is often seen in lupus erythematosus. More rarely, it is also seen in other diseases, such as pellagra, dermatomyositis, and Bloom syndrome.*

Clinician Predicts: *Is it Systemic Lupus Erythematosus?*

[END – success]

E.4. Sample Dialogue from Institutional Knowledge

Sample Dialogue from Institutional Knowledge to Diagnose “Lupus”

Description: Conversation between a clinician and a patient from the Institutional Only baseline, where the patient is suffering from the ground-truth diagnosis, “lupus”. Observe how the patient’s responses tend toward canonical symptoms of lupus, and how each response progressively adds additional details consistent with the diagnosis. This highlights how the Institutional Only baseline may be uniquely informative under our environment \mathcal{E} .

This would explain the primary exception to Finding (2) in Section 5: that although integrating heterogenous information sources often improves overall accuracy, the Institutional Only baseline performs favourably against the multimodal baselines of Random Action selection and Self-Evaluation. Investigating this further, we suggest this is due to the patient response provided by \mathcal{E} being comparatively descriptive, so the institutional knowledge baseline constrains the action space to a uniquely informative subset. It is possible that amending the prompting of \mathcal{E} could mitigate this behaviour.

Clinician Asks: What seems to be the main reason you came to see me today?

Response from \mathcal{E} : I’ve been having a lot of pain and fatigue, and I just don’t feel right.

Clinician Asks: Have you noticed any triggers that make the pain worse?

Response from \mathcal{E} : Sometimes when I’m out in the sun for too long, my joints ache more.

Clinician Asks: Have you experienced any skin rashes or lesions?

Response from \mathcal{E} : Sometimes I get this weird rash on my face, especially after being in the sun. It’s usually on my cheeks and nose.

Clinician Asks: Can you describe the rash on your face in more detail?

Response from \mathcal{E} : It’s a butterfly-shaped rash across my cheeks and nose. It’s red and gets worse when I’m in the sun.

Clinician Asks: Can you tell me more about when your symptoms started and how they have progressed over time?

Response from \mathcal{E} : I’ve been feeling unwell for a few years now. It started with some joint pain and fatigue, and over time I’ve had periods where my skin breaks out in rashes, especially after being in the sun. Sometimes my hands and feet get really swollen. It’s been getting worse, but it’s not always the same.

Clinician Asks: Can you describe any recent changes in your cognitive function or memory?

Response from \mathcal{E} : I’ve been having trouble concentrating lately and sometimes forget things, but it comes and goes.

Clinician Predicts: Is it System Lupus Erythematosus?

[END – success]

E.5. Results for Additional Diagnoses

Below, we present additional results for different ground-truth diagnoses. For each diagnosis, we present a set of violin plots (akin to Figure 2), a set of pie charts (akin to Figure 4), and a sample sequence of actions from CURIOSITREE used to diagnose the condition. We have included a mixture of successful and unsuccessful action sequences in these results.

E.5.1. RESULTS FOR “PRIMARY BILIARY CIRRHOSIS”

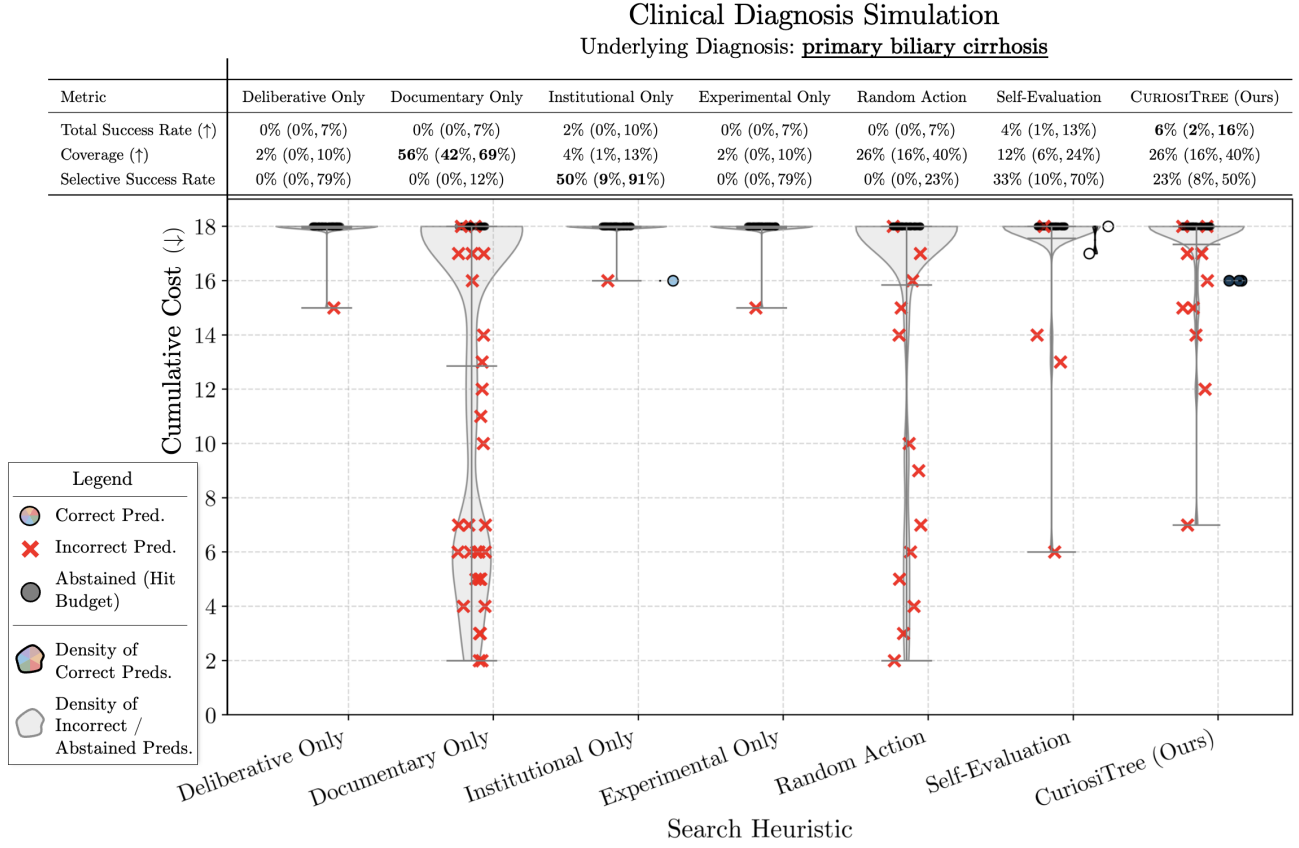


Figure 5: Result of simulation runs with a ground-truth diagnosis “primary biliary cirrhosis”. This figure is interpreted the same way as Figure 2.

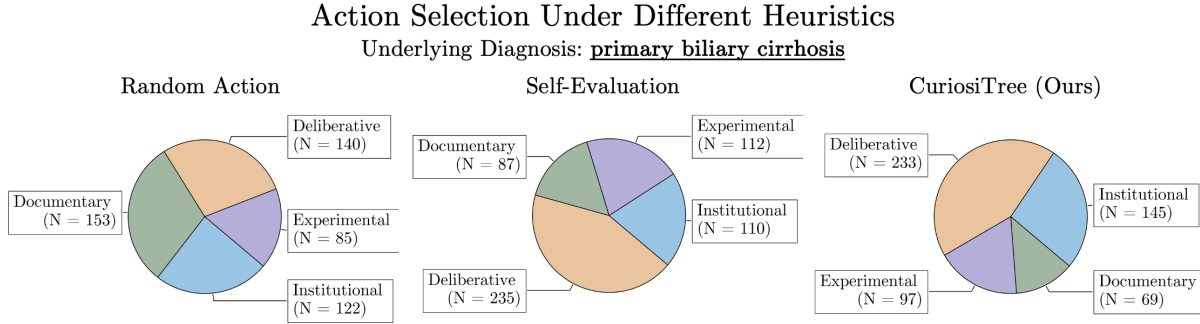


Figure 6: The action classes selected under different policies when diagnosing “primary biliary cirrhosis”. This figure is interpreted the same way as Figure 4.

Attempted Diagnosis of Primary Biliary Cholangitis with CURIOSITREE

Description: A sequence of actions and responses attempting to diagnose “hypertension” using CURIOSITREE. In this case, although the method gets very close—landing on a similar-but-distinct diagnosis, “primary sclerosing cholangitis”—it fails to identify the correct diagnosis.

Clinician Asks: *What seems to be the main reason for your visit today?*

Response from \mathcal{E} : *The patient responds, I've been feeling really tired lately and my skin's been itchy.*

Clinician Reasons: *We know that the patient has been feeling really tired lately and has itchy skin. The patient's symptoms could be indicative of a condition that causes an overproduction of bilirubin or bile salts.*

Response from \mathcal{E} : *N/A*

Clinician Reasons: *We know that the patient has been feeling really tired lately and has itchy skin. This could be indicative of a liver or gallbladder issue that's causing a buildup of bile salts in the skin.*

Response from \mathcal{E} : *N/A*

Clinician Reasons: *We know that the patient has been feeling really tired lately and has itchy skin. The patient's symptoms may be indicative of a condition that affects the flow of bile from the liver to the small intestine.*

Response from \mathcal{E} : *N/A*

Clinician Orders Laboratory Test: *Lipid Profile to evaluate cholesterol and triglyceride levels*

Response from \mathcal{E} : *The test yields:*

- *Total Cholesterol: 320 mg/dL*
- *Triglycerides: 180 mg/dL*
- *LDL Cholesterol: 220 mg/dL*
- *HDL Cholesterol: 40 mg/dL*
- *VLDL Cholesterol: 36 mg/dL*

Clinician Reasons: *We know that the patient has been feeling really tired lately and has itchy skin; this could be indicative of a liver or gallbladder issue that's causing a buildup of bile salts in the skin.*

Response from \mathcal{E} : *N/A*

Clinician Reasons: *We know that the patient's symptoms and lipid profile results are consistent with a condition that affects the flow of bile from the liver to the small intestine; this could be due to a blockage or narrowing of the bile ducts.*

Response from \mathcal{E} : *N/A*

Clinician Orders Laboratory Test: *Endoscopic Retrograde Cholangiopancreatography [ERCP]: Patient's Name: [insert name]; Date of Birth: [insert DOB]; Test: Endoscopic Retrograde Cholangiopancreatography [ERCP]; Reason for Test: Evaluation of bile ducts; Preparation: Patient to fast for 8 hours prior to test; Contrast: ERCP contrast; Tests to Include: Visualization of bile ducts; Collection of bile duct tissue for histology*

Response from \mathcal{E} : *The test yields:*

- *Visualization of bile ducts: Multiple strictures and dilations of intrahepatic bile ducts with beading appearance.*
- *Bile duct histology:*
 - *Portal tracts show chronic inflammation and fibrosis.*
 - *Bile ducts show degenerative changes with bile ductular proliferation.*
 - *No evidence of malignancy.*

Clinician Reasons: *N/A*

Response from \mathcal{E} : *N/A*

Clinician Predicts: *Is it Primary Sclerosing Cholangitis?*

[END – failure]

E.5.2. RESULTS FOR “HYPERTENSION”

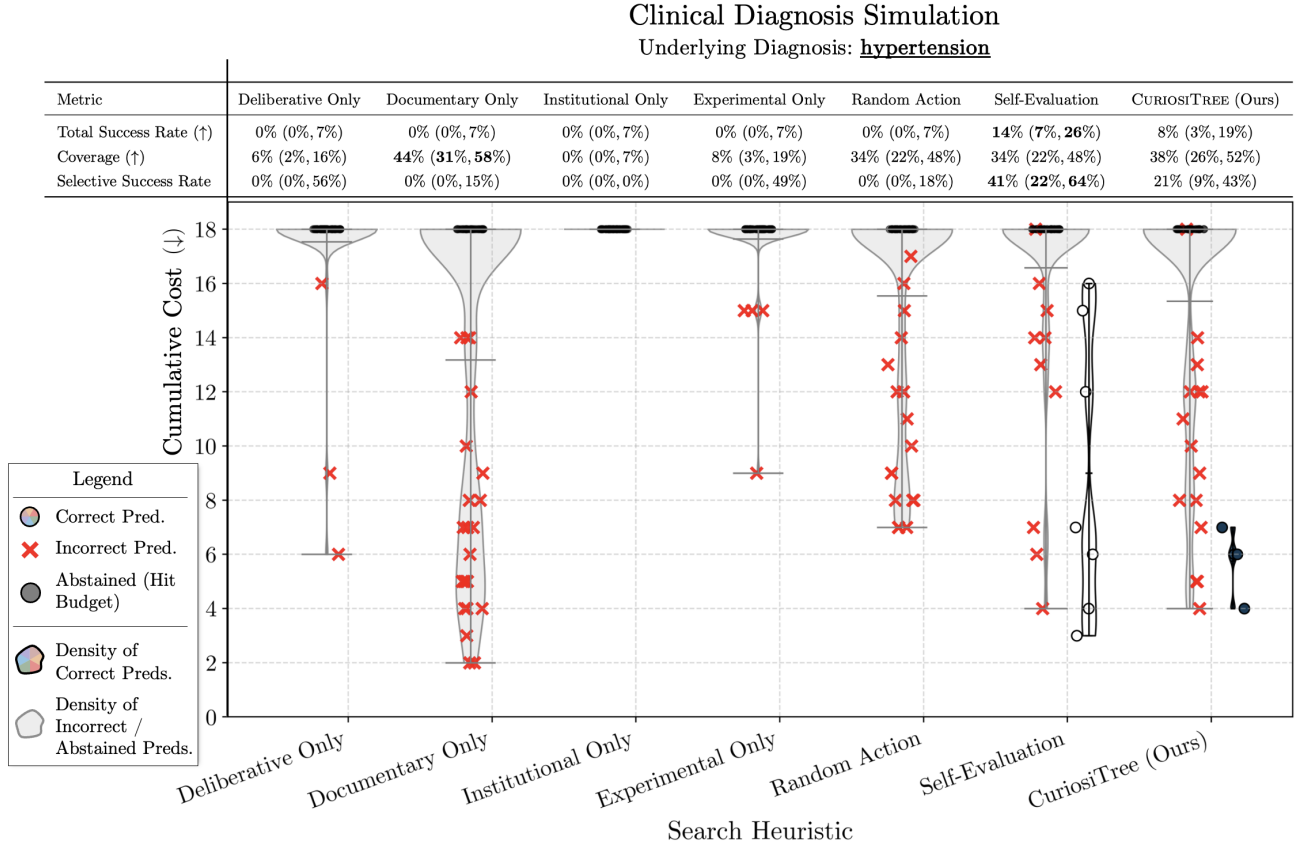


Figure 7: Result of simulation runs with a ground-truth diagnosis “hypertension”. This figure is interpreted the same way as Figure 2.

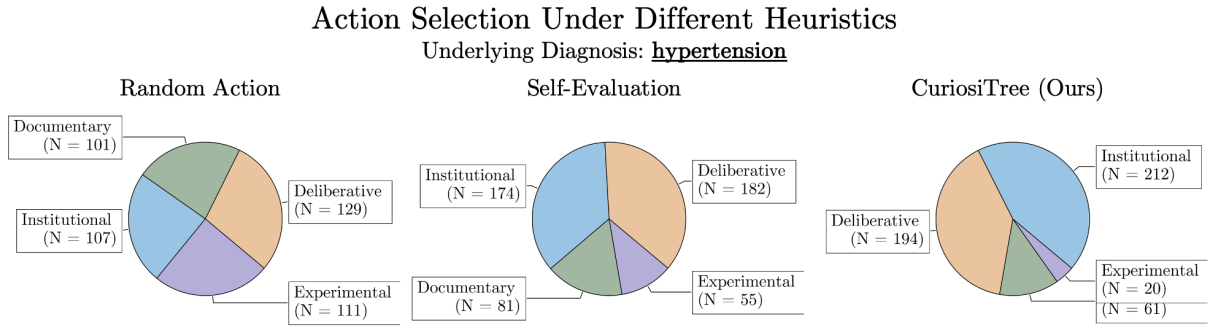


Figure 8: The action classes selected under different policies when diagnosing “hypertension”. This figure is interpreted the same way as Figure 4.

Attempted Diagnosis of “Hypertension” with CURIOSITREE

Description: A sequence of actions and responses attempting to diagnose “hypertension” using CURIOSITREE. In this case, the method fails to remain within the allocated information gathering budget, and so abstains from predicting.

Clinician Asks: *What seems to be the main reason you're visiting the clinic today?*

Response from \mathcal{E} : *The patient responds, "I've been having some trouble lately with headaches, and my family doctor told me to come here for some tests."*

Clinician Asks: *Have you noticed any triggers that seem to cause or worsen your headaches?*

Response from \mathcal{E} : *The patient responds, "Sometimes when I'm stressed or after I've been working outside on a really hot day, my head hurts more."*

Clinician Queries Wikipedia: *Search Query: Heat-related illness symptoms*

Response from \mathcal{E} : *Retrieval: [Heat exhaustion] Heat exhaustion is a heat-related illness characterized by the body's inability to effectively cool itself, typically occurring in high ambient temperatures or during intense physical exertion. In heat exhaustion, core body temperature ranges from 37°C to 40°C (98.6°F to 104°F). Symptoms include profuse sweating, weakness, dizziness, headache, nausea, and lowered blood pressure, resulting from dehydration and serum electrolyte depletion. Heat-related illnesses lie on a spectrum of severity, where heat exhaustion is considered less severe than heat stroke but more severe than heat cramps and heat syncope.*

Clinician Orders Laboratory Test: *Urinalysis: to check for signs of dehydration and assess kidney function*

Response from \mathcal{E} : *The test yields:*

- *Specific gravity: 1.025*
- *pH: 5.5*
- *Protein: 1+*
- *Blood: 0*
- *Glucose: 0*
- *Ketones: 0*
- *WBC: 2–3/hpf*
- *RBC: 0–1/hpf*
- *Bacteria: 0*

Clinician Reasons: *We know that the patient's urinalysis results show a high specific gravity indicating dehydration*

Response from \mathcal{E} : *N/A*

Clinician Asks: *Can you describe any other symptoms you've been experiencing in addition to headaches and what you think might have triggered them?*

Response from \mathcal{E} : *The patient responds, "I've been feeling really tired lately, and my ankles have been a bit swollen. I'm not sure what triggered the headaches, but I did just get back from a big family reunion and we ate a lot of rich foods. I've also been stressed at work, but that's not new."*

Clinician Asks: *Is your urine output less than usual?*

Response from \mathcal{E} : *The patient responds, "No, I think it's about the same."*

Clinician Reasons: *We know that the patient's urine output is normal despite dehydration. This could indicate that the patient's kidneys are still functioning properly.*

Response from \mathcal{E} : *N/A*

Clinician Reasons: *We know that the patient's symptoms are similar to those of heat exhaustion*

Response from \mathcal{E} : *N/A*

Clinician Asks: *Can you describe what you mean by "rich foods" that you ate at the family reunion?*

Response from \mathcal{E} : *The patient responds, I had a lot of fried chicken, mac and cheese, and some potato salad.*

Clinician Reasons: *We know that the patient's symptoms are similar to those of heat exhaustion. This suggests that the patient may be experiencing a heat-related illness.*

Response from \mathcal{E} : *N/A*

E.5.3. RESULTS FOR "TYPE 2 DIABETES MELLITUS"

Clinical Diagnosis Simulation
Underlying Diagnosis: type 2 diabetes mellitus

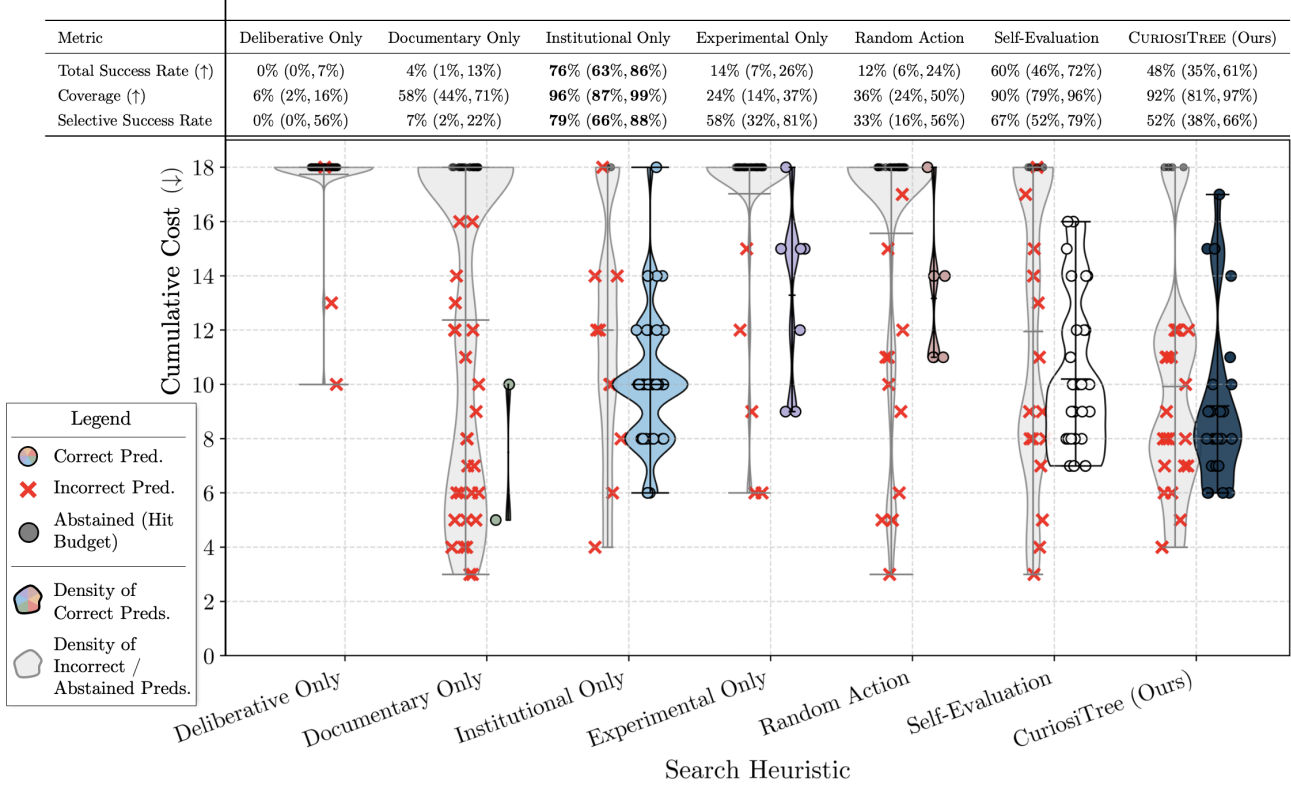


Figure 9: Result of simulation runs with a ground-truth diagnosis "type 2 diabetes mellitus". This figure is interpreted the same way as Figure 2.

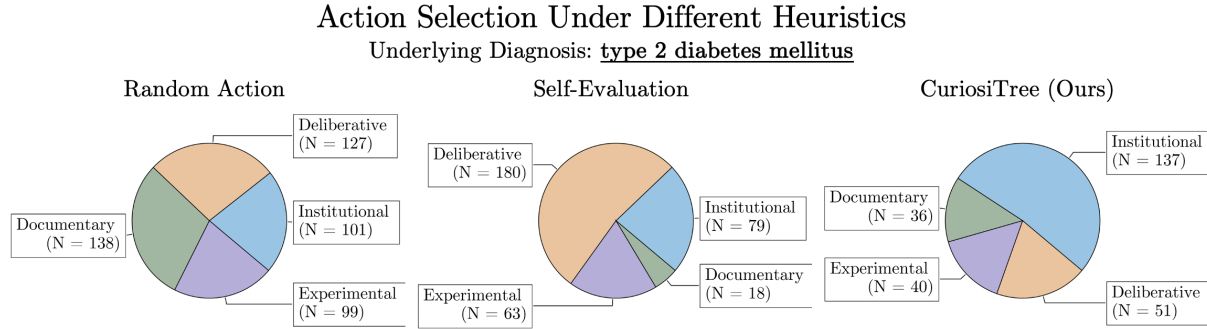


Figure 10: The action classes selected under different policies when diagnosing “type 2 diabetes mellitus”. This figure is interpreted the same way as Figure 4.

Successfully Diagnosing “Diabetes Mellitus Type 2” with CURIOSITREE

Description: A sequence of actions and responses to successfully diagnose “type 2 diabetes mellitus” with CURIOSITREE.

Clinician Asks: *What seems to be the main reason you’ve come to see me today?*

Response from \mathcal{E} : *The patient responds, I’ve been feeling really thirsty lately and I’ve been going to the bathroom a lot.*

Clinician Orders Laboratory Test: *Fasting Plasma Glucose Test: collect a blood sample from the patient after an 8-hour fast to measure glucose levels in the blood*

Response from \mathcal{E} : *The test yields: 182 mg/dL.*

Clinician Predicts: *Is it Diabetes Mellitus Type 2?*

[END – success]

E.5.4. RESULTS FOR “ASTHMA”

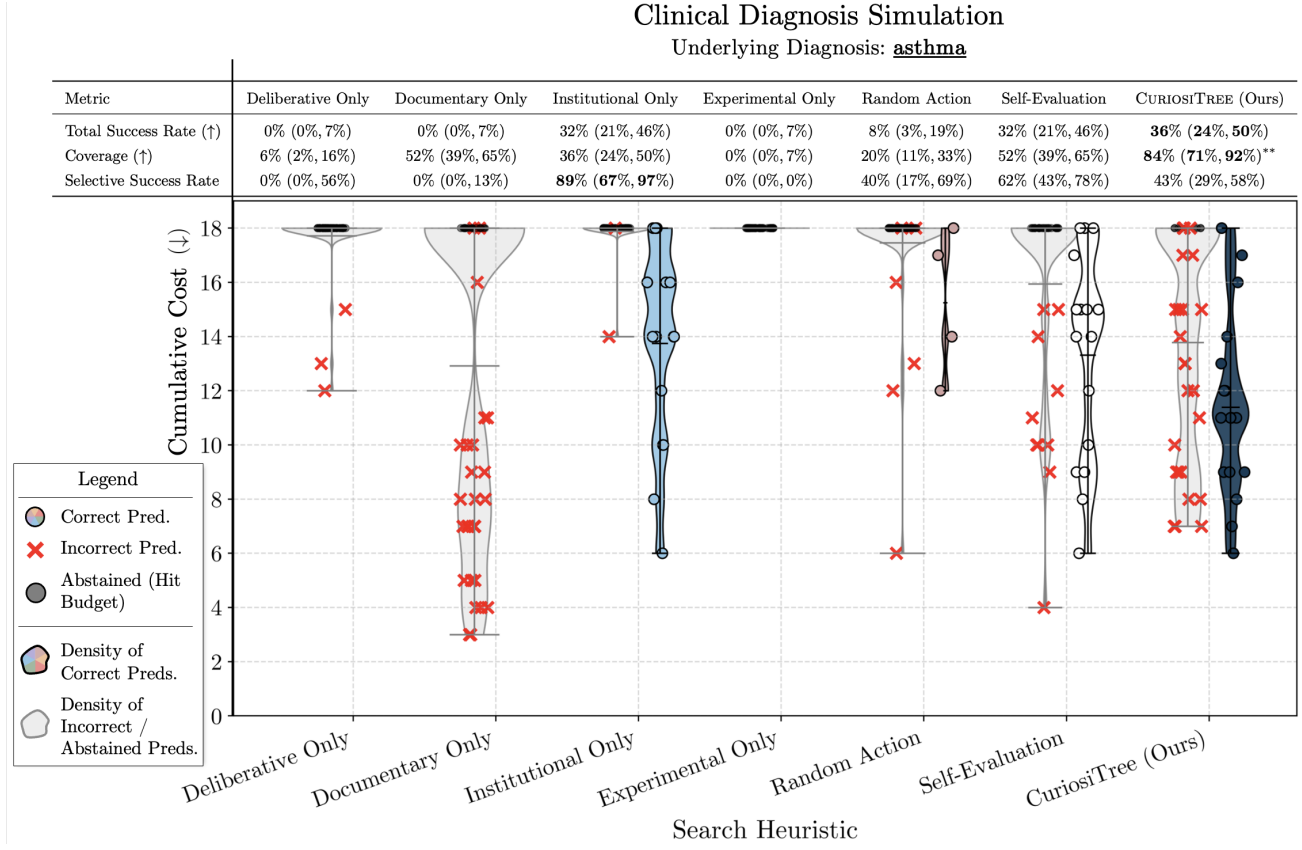


Figure 11: Result of simulation runs with a ground-truth diagnosis “asthma”. This figure is interpreted the same way as Figure 2.

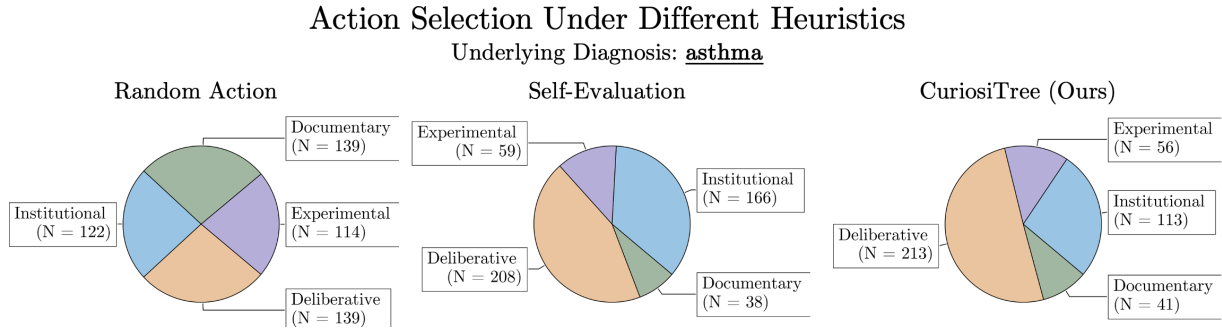


Figure 12: The action classes selected under different policies when diagnosing “asthma”. This figure is interpreted the same way as Figure 4.

Attempted Diagnosis of “Asthma” with CURIOSITREE

Description: A sequence of actions and responses attempting to diagnose “asthma” using CURIOSITREE.

Clinician Asks: *What seems to be the main reason you’re visiting the clinic today?*

Response from \mathcal{E} : *The patient responds, I’ve been having some trouble breathing lately, and it’s been really scary.*

Clinician Orders Laboratory Test: *Pulmonary Function Tests to measure lung function and capacity*

Response from \mathcal{E} : *The test yields:*

- *Forced Expiratory Volume in 1 second (FEV1): 2.1 L (60% predicted)*
- *Forced Vital Capacity (FVC): 3.5 L (80% predicted)*
- *FEV1/FVC ratio: 0.60*
- *Peak Expiratory Flow Rate (PEFR): 250 L/min (50% predicted)*
- *Total Lung Capacity (TLC): 5.2 L (90% predicted)*
- *Residual Volume (RV): 1.8 L (120% predicted)*

Clinician Queries Wikipedia: *Search Query: Chronic obstructive pulmonary disease diagnosis*

Response from \mathcal{E} : *Retrieval: [Chronic obstructive pulmonary disease] Chronic obstructive pulmonary disease (COPD) is a type of progressive lung disease characterized by chronic respiratory symptoms and airflow limitation. GOLD 2024 defined COPD as a heterogeneous lung condition characterized by chronic respiratory symptoms (dyspnea or shortness of breath, cough, sputum production or exacerbations) due to abnormalities of the airways (bronchitis, bronchiolitis) or alveoli (emphysema) that cause persistent, often progressive, airflow obstruction.*

Clinician Predicts: *Is it Chronic Obstructive Pulmonary Disease?*

[END – failure]

E.5.5. RESULTS FOR “RHEUMATOID ARTHRITIS”

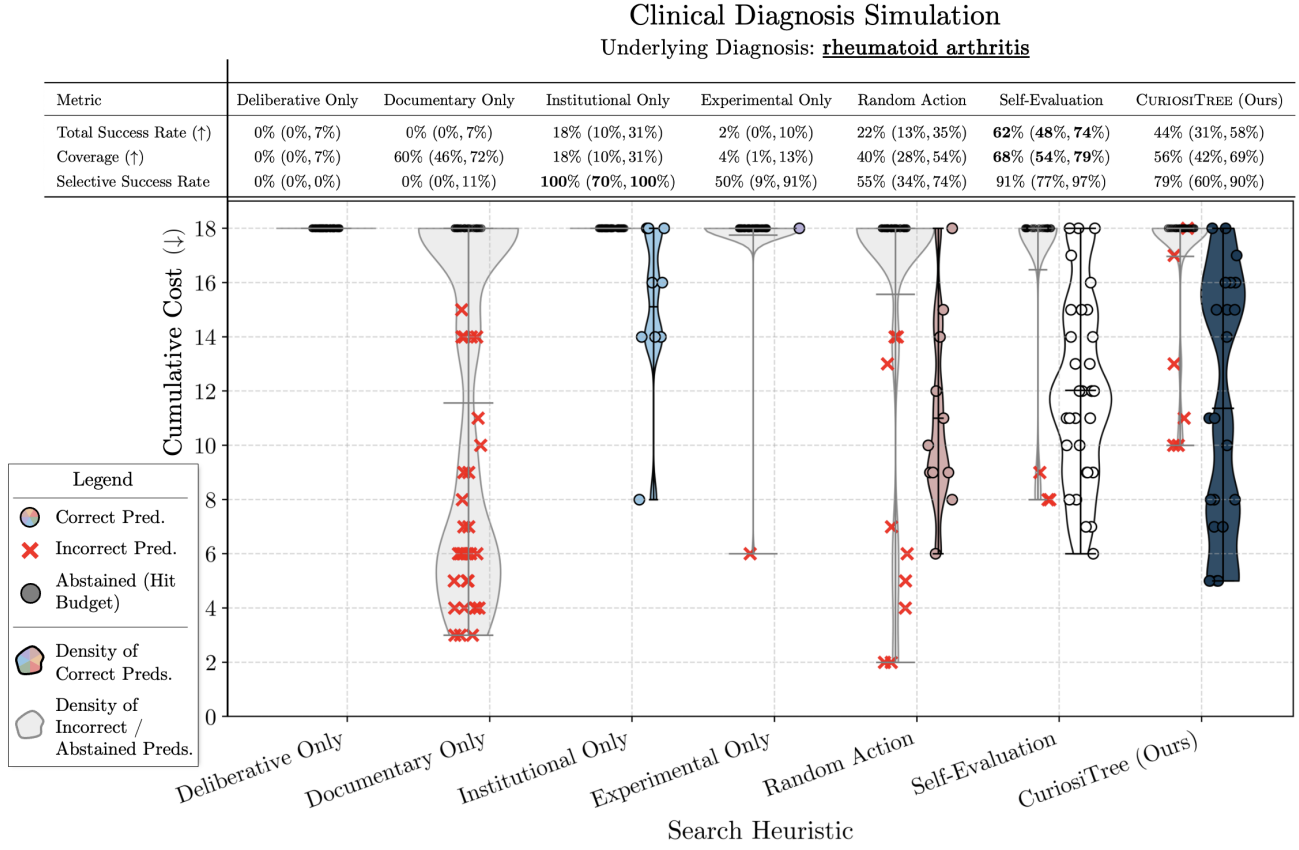


Figure 13: Result of simulation runs with a ground-truth diagnosis “rheumatoid arthritis”. This figure is interpreted the same way as Figure 2.

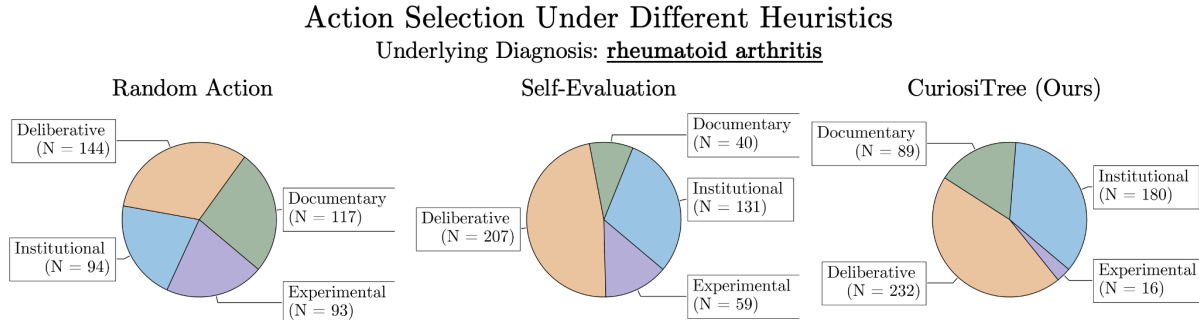


Figure 14: The action classes selected under different policies when diagnosing “rheumatoid arthritis”. This figure is interpreted the same way as Figure 4.

Successfully Diagnosing “Rheumatoid Arthritis” with CURIOSITREE

Description: A sequence of actions and responses leading to the successful diagnosis of “rheumatoid arthritis” using CURIOSITREE.

Clinician Asks: *What seems to be the main reason you're visiting me today?*

Response from \mathcal{E} : *The patient responds, My hands and feet have been hurting a lot lately, and it's been hard to move them in the morning.*

Clinician Reasons: *We know that the patient's symptoms are localized to the hands and feet. This could indicate a condition that primarily affects the distal joints or extremities.*

Response from \mathcal{E} : *N/A*

Clinician Orders Laboratory Test: *Complete Blood Count to evaluate the patient's overall blood health and detect potential signs of inflammation or infection*

Response from \mathcal{E} : *The test yields:*

- *Hemoglobin: 11.8 g/dL*
- *Hematocrit: 36.5%*
- *White Blood Cell Count: 8,500 cells/ μ L*
- *Neutrophils: 6,300 cells/ μ L (74%)*
- *Lymphocytes: 1,800 cells/ μ L (21%)*
- *Monocytes: 400 cells/ μ L (5%)*
- *Eosinophils: 0 cells/ μ L*
- *Basophils: 0 cells/ μ L*
- *Platelet Count: 420,000 cells/ μ L*
- *Mean Corpuscular Volume: 85 fL*
- *Mean Corpuscular Hemoglobin: 28 pg*
- *Mean Corpuscular Hemoglobin Concentration: 33 g/dL*
- *Red Cell Distribution Width: 14.2%*

Clinician Reasons: *We know that the patient's hemoglobin is slightly below the normal range. This mild anemia could be a sign of chronic inflammation or another underlying condition.*

Response from \mathcal{E} : *N/A*

Clinician Asks: *Have you noticed any skin rashes or lesions on your hands or feet?*

Response from \mathcal{E} : *The patient responds, "I've had some weird bumps on my feet, but I don't think it's a big deal."*

Clinician Asks: *Can you describe the bumps on your feet in more detail?*

Response from \mathcal{E} : *The patient responds, They're painful and swollen, usually on the balls of my feet and toes.*

Clinician Orders Laboratory Test: *X-ray of Hands and Feet: Patient ID: [insert ID number]; Test Name: X-ray of Hands and Feet; Test Type: Imaging; Sample Type: N/A; Collection Method: Radiography; Special Instructions: Patient to remove jewelry and clothing from hands and feet; Clinical Information: Patient presenting with joint pain and swelling in hands and feet*

Response from \mathcal{E} : *The test yields: **Hands:***

- *Bilateral joint space narrowing of the 2nd and 3rd metacarpophalangeal joints*
- *Periarticular erosions in the 2nd proximal interphalangeal joint*

- Soft tissue swelling around the 3rd metacarpophalangeal joint

Feet:

- Joint space narrowing of the 1st metatarsophalangeal joint
- Periarticular erosions in the 2nd metatarsophalangeal joint
- Bilateral hallux valgus deformity

Clinician Reasons: This suggests a condition that primarily affects the distal joints or extremities and is likely inflammatory in nature

Response from \mathcal{E} : N/A

Clinician Predicts: Is it Rheumatoid Arthritis?

[END – success]

E.5.6. RESULTS FOR “MULTIPLE SCLEROSIS”

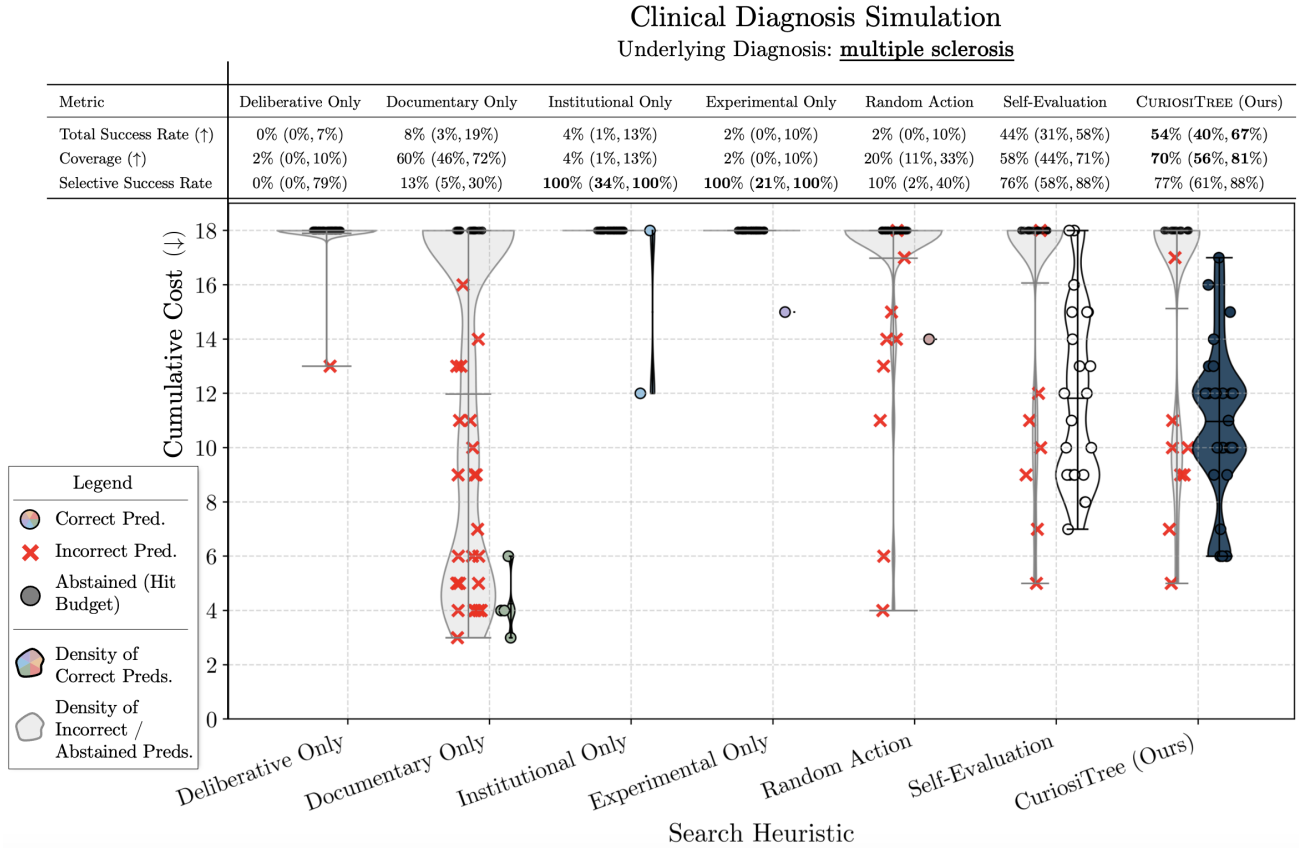


Figure 15: Result of simulation runs with a ground-truth diagnosis “multiple sclerosis”. This figure is interpreted the same way as Figure 2.

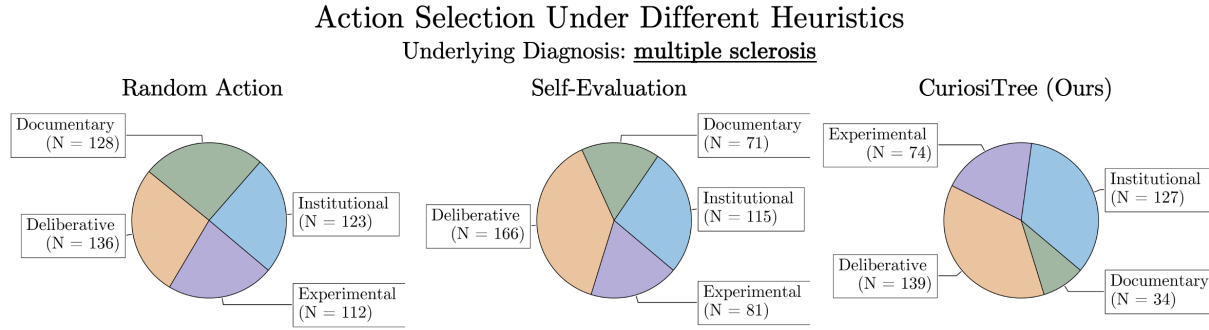


Figure 16: The action classes selected under different policies when diagnosing “multiple sclerosis”. This figure is interpreted the same way as Figure 4.

Successfully Diagnosing “Multiple Sclerosis” with CURIOSITREE

Description: A sequence of actions and responses leading to the successful diagnosis of “multiple sclerosis” using CURIOSITREE.

Clinician Asks: *What seems to be the main reason you’ve come to see me today?*

Response from \mathcal{E} : *The patient responds, I’ve been having some trouble with my vision and balance, and I’ve been feeling really tired lately.*

Clinician Orders Laboratory Test: *Magnetic Resonance Imaging of the Brain to evaluate for structural abnormalities or lesions affecting vision and balance*

Response from \mathcal{E} : *The test yields:*

- *T1-weighted images show multiple periventricular and subcortical white matter lesions, with 5 lesions demonstrating enhancement after gadolinium administration.*
- *T2-weighted images reveal a total of 12 hyperintense lesions in the periventricular, subcortical, and infratentorial regions.*
- *FLAIR sequence shows multiple hyperintense lesions in the supratentorial and infratentorial regions, with involvement of the optic radiations and the middle cerebellar peduncles.*
- *No evidence of acute hemorrhage, mass effect, or significant atrophy.*

Clinician Reasons: *We know that the patient’s symptoms and MRI findings are not suggestive of an acute or infectious process. The presence of enhancing and non-enhancing lesions suggests a chronic and relapsing condition such as multiple sclerosis.*

Response from \mathcal{E} : *N/A*

Clinician Reasons: *We know that the presence of both enhancing and non-enhancing lesions on the patient’s MRI suggests a chronic and relapsing condition*

Response from \mathcal{E} : *N/A*

Clinician Predicts: *Is it Multiple Sclerosis?*

[END – success]

E.5.7. RESULTS FOR “CYSTIC FIBROSIS”

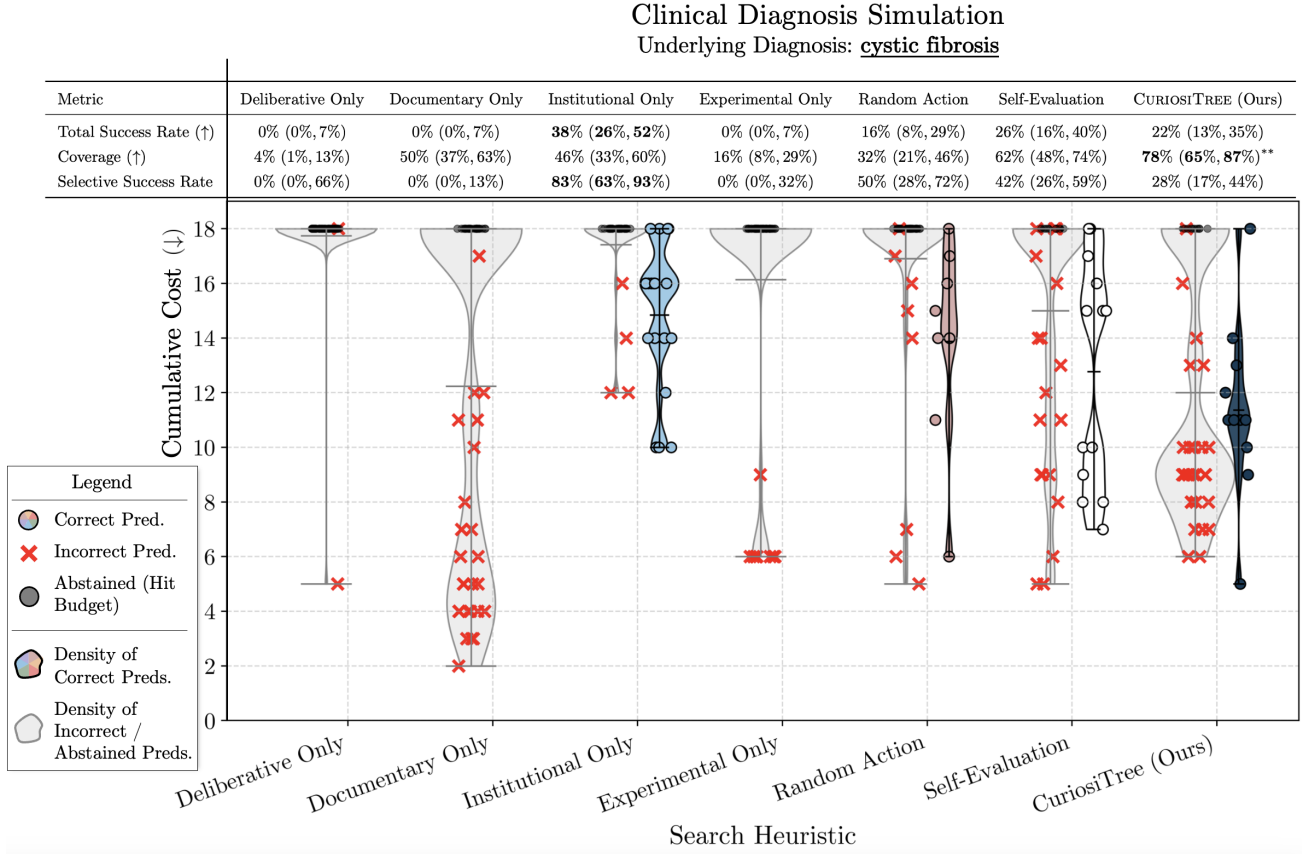


Figure 17: Result of simulation runs with a ground-truth diagnosis “cystic fibrosis”. This figure is interpreted the same way as Figure 2.

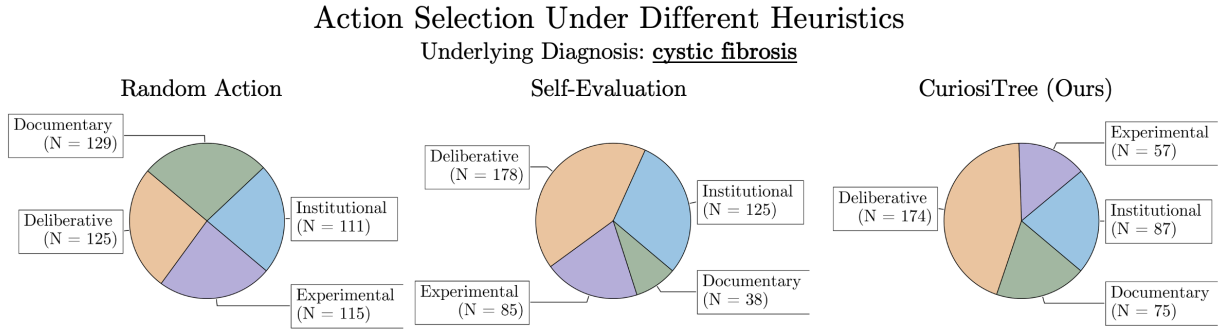


Figure 18: The action classes selected under different policies when diagnosing “cystic fibrosis”. This figure is interpreted the same way as Figure 4.

Unsuccessfully Diagnosing “Cystic Fibrosis” with CURIOSITREE

Description: A sequence of actions and responses attempting to diagnose “cystic fibrosis” using CURIOSITREE.

Clinician Asks: *What brings you to the clinic today?*

Response from \mathcal{E} : *The patient responds, I’ve been having some trouble breathing and I’ve had a lot of lung*

infections lately.

Clinician Queries Wikipedia: Search Query: Causes of recurrent lung infections

Response from \mathcal{E} : Retrieval: [Post-tuberculosis lung disease] Post-tuberculosis lung disease (PTLD) is ongoing lung disease that is caused by tuberculosis (TB) but persists after the infection is cured. PTLD can affect the airways, lung parenchyma, pulmonary vasculature, and pleura.

Clinician Orders Laboratory Test: Pulmonary Function Test to assess lung function and detect any obstructive or restrictive patterns

Response from \mathcal{E} : The test yields:

- Forced Expiratory Volume in 1 second (FEV1): 1.8 L (55% predicted)
- Forced Vital Capacity (FVC): 2.5 L (65% predicted)
- FEV1/FVC ratio: 0.72
- Residual Volume (RV): 1.2 L (120% predicted)
- Total Lung Capacity (TLC): 4.2 L (85% predicted)

Clinician Queries Wikipedia: Search Query: Chronic lung disease causes and symptoms

Response from \mathcal{E} : Retrieval: [Chronic obstructive pulmonary disease] Chronic obstructive pulmonary disease (COPD) is a type of progressive lung disease characterized by chronic respiratory symptoms and airflow limitation. GOLD 2024 defines COPD as a heterogeneous lung condition characterized by chronic respiratory symptoms (dyspnea, cough, sputum production, or exacerbations) due to abnormalities of the airways (bronchitis, bronchiolitis) or alveoli (emphysema) that cause persistent, often progressive, airflow obstruction.

Clinician Reasons: We know that the patient's Pulmonary Function Test results show a decreased FEV1/FVC ratio and increased Residual Volume. These findings are consistent with a diagnosis of chronic obstructive pulmonary disease.

Response from \mathcal{E} : N/A

Clinician Predicts: Is it Chronic obstructive pulmonary disease?

[END – failure]

E.5.8. RESULTS FOR “AMYOTROPHIC LATERAL SCLEROSIS”

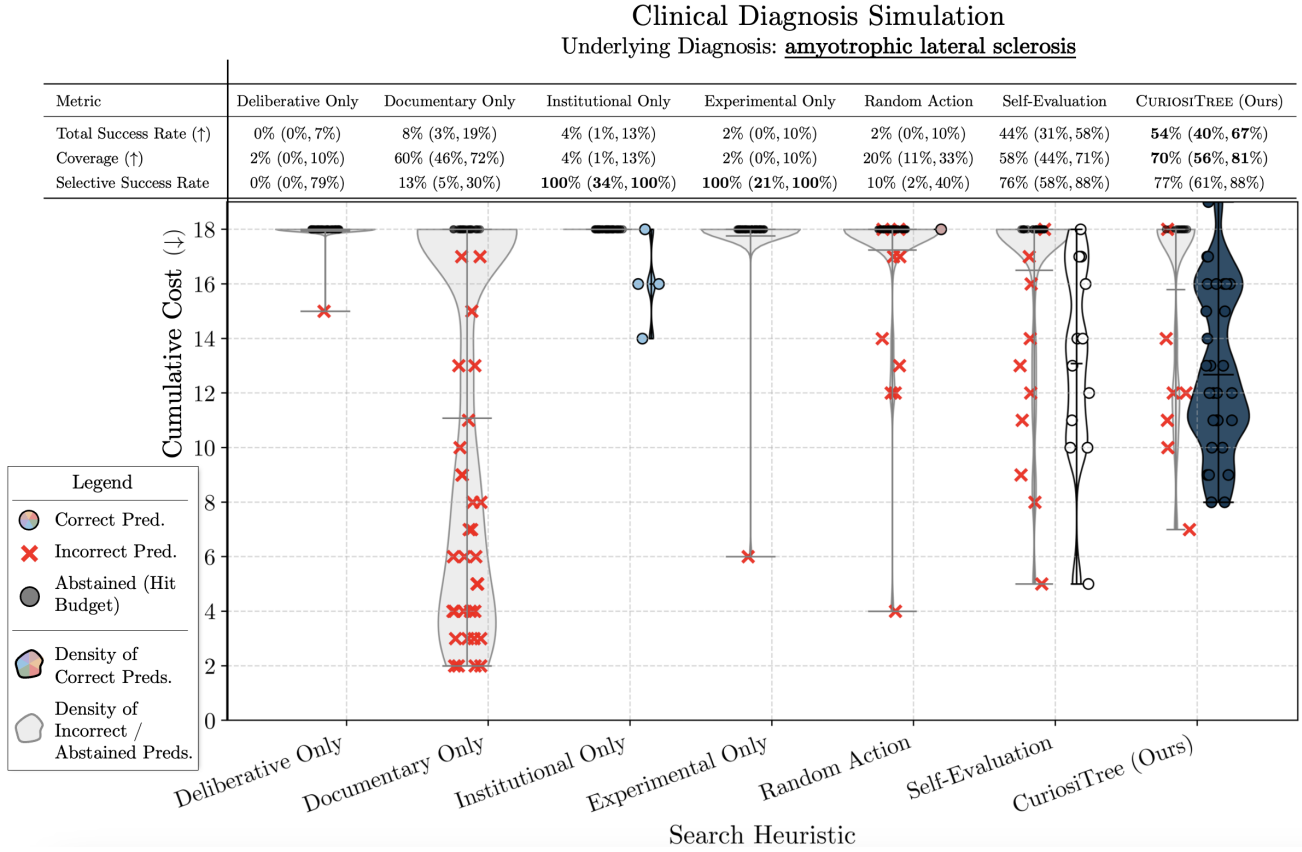


Figure 19: Result of simulation runs with a ground-truth diagnosis “amyotrophic lateral sclerosis”. This figure is interpreted the same way as Figure 2.

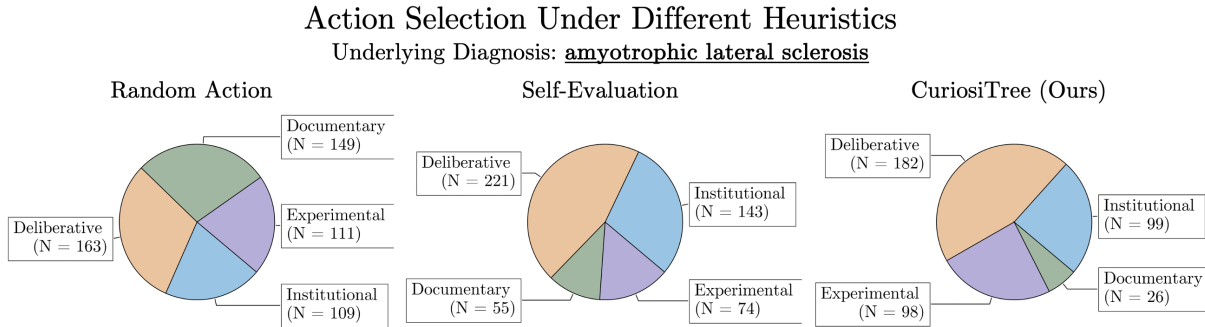


Figure 20: The action classes selected under different policies when diagnosing “amyotrophic lateral sclerosis”. This figure is interpreted the same way as Figure 4.

Successfully Diagnosing “Amyotrophic Lateral Sclerosis” with CURIOSITREE

Description: A sequence of actions and responses leading to the successful diagnosis of “amyotrophic lateral sclerosis” using CURIOSITREE.

Clinician Asks: *What seems to be the main reason you're visiting the clinic today?*

Response from \mathcal{E} : *The patient responds, I've been having some trouble with my arm, it's been feeling weak and I'm having a hard time holding things.*

Clinician Orders Laboratory Test: *Creatine Kinase level to evaluate muscle damage or disease*

Response from \mathcal{E} : *The test yields: 543 U/L.*

Clinician Orders Laboratory Test: *Muscle Biopsy to examine muscle tissue for signs of disease or damage*

Response from \mathcal{E} : *The test yields: Muscle biopsy shows denervation atrophy, with groups of atrophic fibers and scattered angulated fibers, and evidence of reinnervation, including fiber type grouping.*

Clinician Reasons: *which is consistent with a motor neuron disease or peripheral neuropathy*

Response from \mathcal{E} : *N/A*

Clinician Reasons: *We know that the patient's elevated creatine kinase level and muscle biopsy results suggest an ongoing process of muscle damage and repair*

Response from \mathcal{E} : *N/A*

Clinician Reasons: *We know that the patient's elevated creatine kinase level and muscle biopsy results suggest an ongoing process of muscle damage and repair*

Response from \mathcal{E} : *N/A*

Clinician Reasons: *We know that the combination of muscle weakness and denervation atrophy suggests a disease affecting the motor neurons*

Response from \mathcal{E} : *N/A*

Clinician Asks: *Do you have difficulty swallowing or speaking?*

Response from \mathcal{E} : *The patient responds, "Sometimes I have trouble getting words out, and swallowing food can be a bit of a problem, especially with thicker foods."*

Clinician Reasons: *We know that the patient's symptoms and laboratory results are suggestive of a specific type of motor neuron disease that affects both upper and lower motor neurons*

Response from \mathcal{E} : *N/A*

Clinician Predicts: *Is it Amyotrophic Lateral Sclerosis?*

[END – success]

E.5.9. RESULTS FOR “MYASTHENIA GRAVIS”

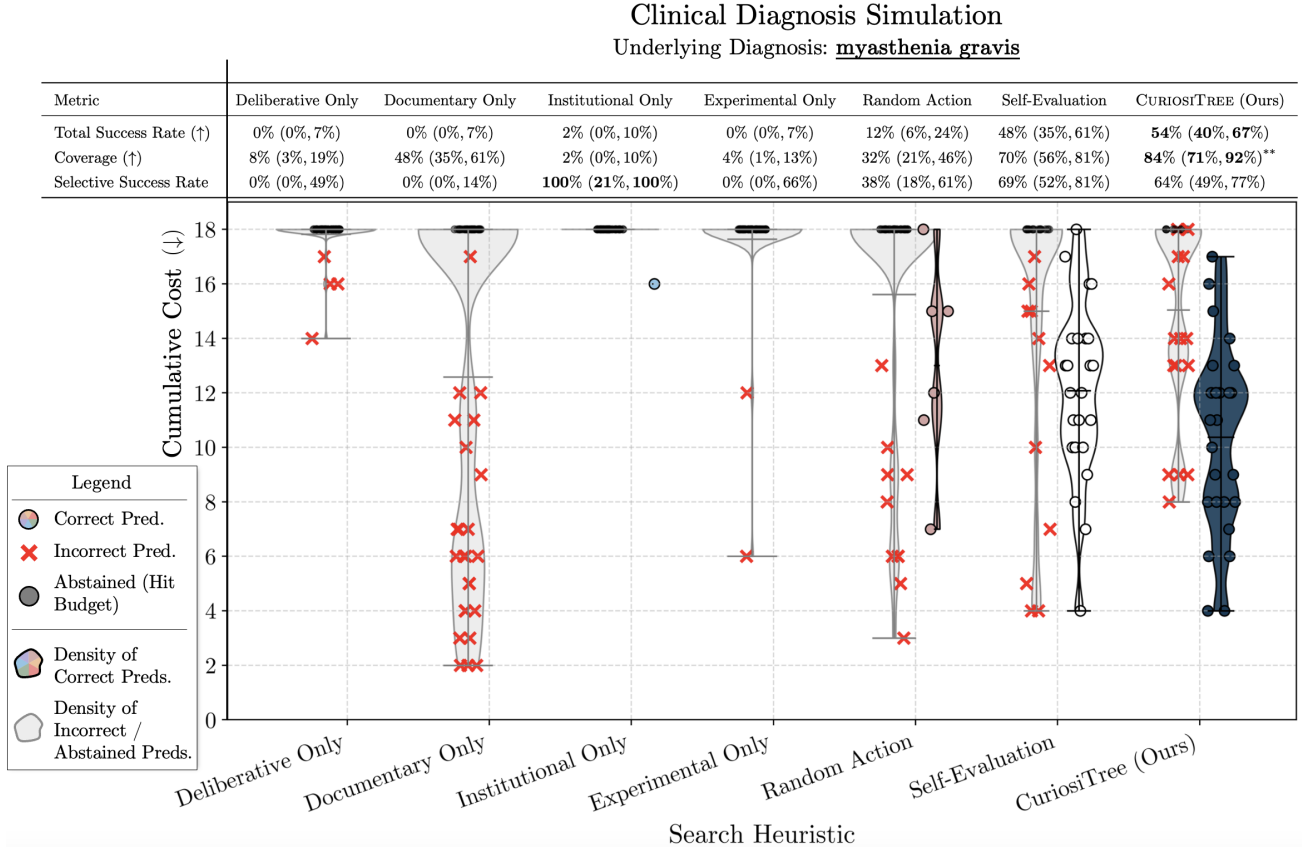


Figure 21: Result of simulation runs with a ground-truth diagnosis “myasthenia gravis”. This figure is interpreted the same way as Figure 2.

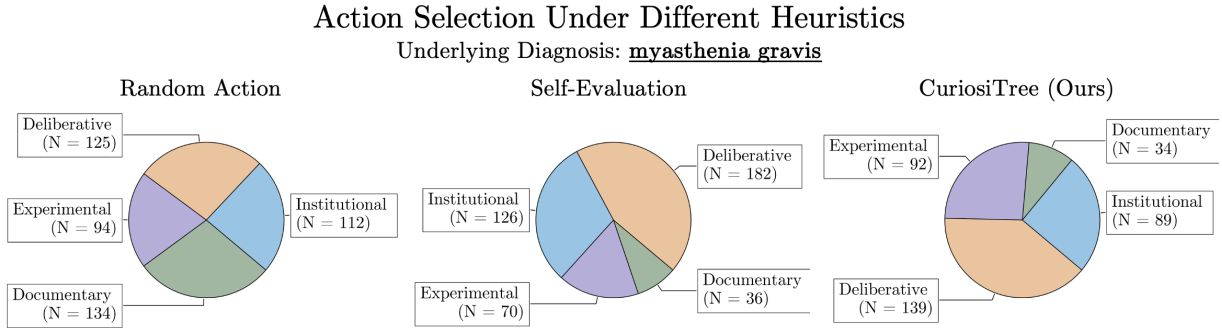


Figure 22: The action classes selected under different policies when diagnosing “myasthenia gravis”. This figure is interpreted the same way as Figure 4.

Successfully Diagnosing “Myasthenia Gravis” with CURIOSITREE

Description: A sequence of actions and responses leading to the successful diagnosis of “myasthenia gravis” using CURIOSITREE.

Clinician Asks: *What brings you to the clinic today?*

Response from \mathcal{E} : *The patient responds, I've been having some trouble with my muscles feeling weak, especially after I've been active or at the end of the day.*

Clinician Orders Laboratory Test: *Initialized by: Dr. [Last Name] Date: [Current Date] Requested Test: Electromyography to assess the electrical activity of the muscles and diagnose conditions such as myasthenia gravis or amyotrophic lateral sclerosis*

Response from \mathcal{E} : *The test yields:*

- *Decreased compound muscle action potential (CMAP) amplitudes with repetitive nerve stimulation (RNS)*
- *A 30% decrement in CMAP amplitude at 3 Hz stimulation of the ulnar nerve*
- *Single-fiber electromyography (SFEMG) showing increased jitter (mean consecutive difference, MCD: 55 μ s) and blocking (11% of fibers)*

Clinician Predicts: *Is it Myasthenia Gravis?*

[END – success]