Timely Clinical Diagnosis through Active Test Selection

Anonymous Author(s)
Affiliation
Address
email

Abstract

There is growing interest in using machine learning (ML) to support clinical diagnosis, but most approaches rely on static, fully observed datasets and fail to reflect the sequential, resource-aware reasoning clinicians use in practice. Diagnosis remains complex and error prone, especially in high-pressure or resource-limited settings, underscoring the need for frameworks that help clinicians make timely and cost-effective decisions. We propose ACTMED (Adaptive Clinical Test selection via Model-based Experimental Design), a diagnostic framework that integrates Bayesian Experimental Design (BED) with large language models (LLMs) to better emulate real-world diagnostic reasoning. At each step, ACTMED selects the test expected to yield the greatest reduction in diagnostic uncertainty for a given patient. LLMs act as flexible simulators, generating plausible patient state distributions and supporting belief updates without requiring structured, task-specific training data. Clinicians can remain in the loop; reviewing test suggestions, interpreting intermediate outputs, and applying clinical judgment throughout. We evaluate ACTMED on real-world datasets and show it can optimize test selection to improve diagnostic accuracy, interpretability, and resource use. This represents a step toward transparent, adaptive, and clinician-aligned diagnostic systems that generalize across settings with reduced reliance on domain-specific data.

1 Introduction

Clinical diagnosis is a fundamental step in modern medical practice [1], providing the framework for future investigations and guiding treatment decisions, often determining patient outcomes. Yet it remains complex and error-prone, especially in fast-paced or resource-limited settings [2], where delays, misdiagnoses, and over-testing pose persistent global challenges [3], [4]. Additionally, the WHO projects a global shortage of more than 12 million qualified health professionals by 2035 [5]. Machine learning (ML) has emerged as a promising tool to support clinicians by improving diagnostic accuracy, optimizing test selection, and enabling earlier disease detection [6]–[9]. However, many current ML models operate under unrealistic assumptions, such as complete data availability [10], and fail to reflect the iterative, context-aware decision-making used by human clinicians [11].

Clinical diagnosis. Clinical diagnosis has traditionally followed local or national guidelines based on clinical trials and expert consensus [12], [13]. Although such guidelines improve outcomes by standardizing care, they are population-based and often inefficient at the individual level, with an estimated 40 to 60% of diagnostic tests being unnecessary [14]. Resource constraints further hinder access; for example, around 15% of clinician-ordered genetic tests go unperformed due to financial barriers [15]. In response, machine learning (ML) models have been proposed to support more personalized and efficient test ordering [16]. However, for these models to gain trust and adoption, they must be transparent and aligned with clinicians' reasoning processes [17], [18].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Current models for diagnosis. The diagnostic process is inherently sequential, involving stepwise information gathering from patient examinations and tests [19]. Clinicians aim to achieve accurate early diagnoses while minimizing diagnostic costs, as timely intervention can significantly improve outcomes [20]–[22]. Prior models for early diagnosis often target a single disease and rely on specific modalities such as blood tests or imaging [23]–[25], which typically require large training data sets and assume full availability of modality [26]. In reality, this assumption rarely holds, and while imputation methods can handle missing data, they often introduce bias [10]. Additionally, many models treat diagnosis as a static classification task, overlooking the inherently dynamic and progressive nature of disease development and clinical reasoning. State-space models have been developed to capture disease trajectories [27]–[29], but they also require extensive retraining and struggle with balancing information acquisition costs [30]. Consequently, there remains a gap for generalizable frameworks that can reason dynamically under uncertainty and resource constraints [22], [31].

LLMs for clinical diagnosis. Recent advances in large language models (LLMs) have sparked interest in their use as general-purpose tools for medical decision-making [32], [33]. LLMs perform well on medical licensing exams [18], [34], [35] and are particularly effective in zero-shot settings [36], [37]. However, their direct deployment in clinical contexts faces challenges, including limited transparency and interpretability [38]. While chain-of-thought prompting improves interpretability [39], LLMs often deviate from the probabilistic optimum, although fine-tuning can improve their probabilistic reasoning [40]. Furthermore, recent work shows that LLM explanations often do not reflect their true internal reasoning processes, raising concerns about the faithfulness of chain-of-thought outputs [41]. It has also been shown that LLMs can approximate structured decision-making tasks, such as Bayesian optimization or decision tree induction, by leveraging latent inductive biases learned from large-scale text corpora [42]–[44]. Additionally, shifting LLM reasoning to the natural language solution space has been shown to enhance decision quality [45].

Contributions. ① We motivate and formalize a transparent, stepwise diagnostic framework that aligns with clinical reasoning. ② We propose ACTMED, a probabilistic approach to timely diagnosis that uses Bayesian Experimental Design with LLMs to adaptively select tests based on their expected diagnostic utility. ③ We show that shifting reasoning from the LLM to the natural language output space can improve clinical decision-making. ④ We validate ACTMED on real-world datasets, demonstrating its ability to optimize test selection and improve diagnostic accuracy, interpretability, and resource use. This framework ultimately contributes to more transparent, adaptive, and clinician-aligned diagnostic processes.

2 Problem formalism

Agent-based diagnosis model. Let $T = \{1, 2, \dots, T_{\max}\}$ denote the discrete time horizon representing stages in the decision process. The space of natural language is denoted by Σ, and $\mathcal{S} \in \Sigma$ represents the natural language instructions provided to the agent at each stage. The agent must return a diagnosis $d_t \in \mathcal{D}_t \subset \Sigma$, where $\mathcal{D}_t$ is the set of possible diagnoses at time t. We assume that a patient may present with a subset of all possible diagnoses $\mathcal{D}_{\text{true},t} \subset \mathcal{D}_t$. The agent independently estimates the posterior probability $P(y_{d_t} = 1 \mid K_t)$ for each diagnosis $d_t \in \mathcal{D}_t$, where K_t is the information available at time t, and the belief is updated dynamically as new information is acquired.

At each time step $t \in T$, the agent observes a subset $K_t \subset \mathcal{X}_t$ of ground truth information $\mathcal{X}_t \subset \Sigma$, and may request additional information $u'_t \in \mathcal{U}_t$ from an external source, with $\mathcal{U}_t = \mathcal{X}_t \setminus K_t$. The agent's objective is to minimize the diagnostic error and cumulative cost of information:

$$\min_{\{\hat{y}_{d_t}\}_{d_t \in D_t}} \sum_{t \in T} \mathbb{I}[\hat{y}_{d_t} \neq y_{d_t}], \qquad \min_{\{u'_t\}_{t \in T}} \sum_{t \in T} c(u'_t), \tag{1}$$

where $c(u'_t)$ is the cost of requested information and $\mathbb{I}[\hat{y}_{d_t} \neq y_{d_t}]$ is the 0-1 loss for an incorrect diagnosis at decision point d_t.

Optimal diagnostic test selection. The challenge of optimal diagnostic test selection within Bayesian Experimental Design (BED) is to identify the test, $u_t^{(i)}$, that provides the greatest informational utility regarding a diagnostic label, y_{d_t}. We consider a binary classification (e.g., sick/not sick), though this framework extends to multiple diagnoses by independent simultaneous application. The

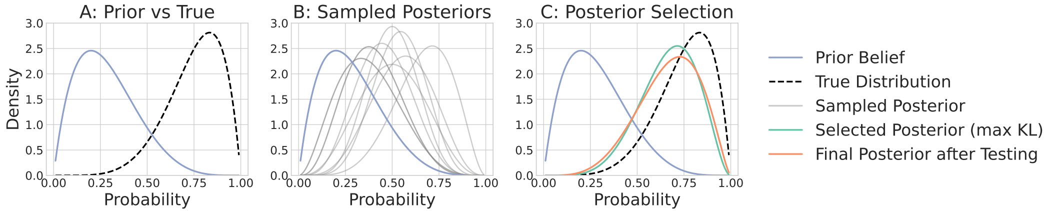

Figure 1: *Illustration of KL divergence-based diagnostic test selection.* **(A)** The agent starts with a prior belief (blue) and aims to approximate the true posterior (black dashed). **(B)** It simulates hypothetical posteriors (gray) for candidate tests and computes KL divergence from the prior. **(C)** The test yielding the highest divergence (green) is selected. After observing the result, the belief is updated to the final posterior (orange).

---

core objective is to select tests that maximally reduce epistemic uncertainty, the uncertainty stemming from our limited knowledge or model imperfections, which is reducible with new data, as opposed to aleatoric uncertainty, which is inherent system randomness.

To model the impact of potential information $u_t^{(i)}$, we employ a surrogate model to draw M hypothetical outcomes $u_t^{(i,j)} \sim P(u_t^i)$. For binary classification, we assume the posterior distribution $P(y_{d_t} = 1 \mid K_t, u_t^{(i)})$ follows a Bernoulli distribution $\mathbb{B}(p_i)$, where $p_i \in [0, 1]$ is the success probability. In clinical practice, a test's value lies not just in reducing uncertainty, but in meaningfully shifting the probability of disease presence, especially across decision thresholds relevant to treatment decisions. While entropy-based formulations can be used, they may sometimes prioritize tests that reinforce confident but incorrect predictions. For instance, when a patient is initially assigned a very low disease probability, a truly informative test may increase this belief substantially. However, a test with no real diagnostic value might keep the prediction near zero, deceptively minimizing entropy (see Appendix B).

The information gain can also be expressed as the difference in KL divergence between the posterior and prior distributions of y_{d_t}. Maximizing this expectation ensures the selection of tests whose outcomes, on average, induce the most significant and diagnostically meaningful shifts in belief. This aligns with the core BED principle of maximizing information gain and directly addresses the clinical need to understand how a test will alter diagnostic probabilities, especially concerning critical decision thresholds. Given the unknown nature of these prior and posterior distributions, we utilize our surrogate model to generate samples j representing hypothetical test results. Let p_{prior} represent the prior, and p_{post} the posterior probability distribution: $p_{\text{prior}}^{(j)} \sim P(y_{d_t} = 1 \mid K_t)$, $\quad p_{\text{post}}^{(j)} \sim P(y_{d_t} = 1 \mid K_t, u_t^{(i,j)})$. The expected KL divergence is then computed as the average KL divergence over the M samples from both distributions:

$$\mathbb{E}[\text{KL}(\mathbb{B}(p_{\text{post}}) \parallel \mathbb{B}(p_{\text{prior}}))] = \frac{1}{M} \sum_{j=1}^{M} p_{\text{post}}^{(j)} \log \left(\frac{p_{\text{post}}^{(j)}}{p_{\text{prior}}^{(j)}} \right) + \left(1 - p_{\text{post}}^{(j)} \right) \log \left(\frac{1 - p_{\text{post}}^{(j)}}{1 - p_{\text{prior}}^{(j)}} \right). \tag{2}$$

The optimal piece of information $u_t^{(i^*)}$ is the one that maximizes this expected KL divergence $i^* = \arg\max_i \, \mathbb{E}[\text{KL}(\mathbb{B}(p_{\text{post}}) \parallel \mathbb{B}(p_{\text{prior}}))]$. The process of KL-guided diagnostic test selection and belief updating is illustrated in Figure 1.

Example 1: An agent is tasked with diagnosing chronic kidney disease (CKD), aiming to establish the correct diagnosis d_t as early as possible while minimizing additional diagnostic evaluations u_t^i due to budget constraints and test delays. At each time point t, the agent has access to clinical information K_t, including demographics and previous test results, and can request further information u_t^i, such as lab tests or imaging, to refine the diagnosis d_t.

3 ACTMED: Probabilistic reasoning for clinical diagnosis

Surrogate model sampling. ACTMED relies on evaluating the expected utility of potential diagnostic tests by estimating hypothetical posteriors over diagnoses. Specifically, for each candidate

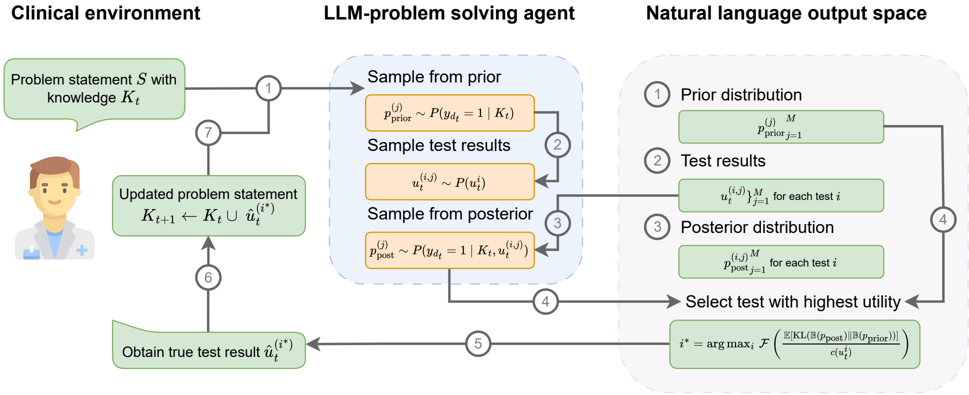
Clinical environment
LLM-problem solving agent
Natural language output space
Problem statement S with knowledge K_t
1
7
Updated problem statement $K_{t+1} \leftarrow K_t \cup \hat{u}_t^{(i^*)}$
6
Obtain true test result $\hat{u}_t^{(i^*)}$
Sample from prior
$p_{\text{prior}}^{(j)} \sim P(y_{d_t} = 1 \mid K_t)$
2
Sample test results
$u_t^{(i,j)} \sim P(u_t^i)$
Sample from posterior
3
$p_{\text{post}}^{(j)} \sim P(y_{d_t} = 1 \mid K_t, u_t^{(i,j)})$
4
5
1 Prior distribution
${p_{\text{prior}}^{(j)}}_{j=1}^{M}$
2 Test results
$u_t^{(i,j)}\}_{j=1}^{M}$ for each test i
3 Posterior distribution
${p_{\text{post}}^{(i,j)}}_{j=1}^{M}$ for each test i
4
Select test with highest utility
$i^* = \arg\max_i \mathcal{F}\left(\frac{\mathbb{E}[\mathrm{KL}(\mathbb{B}(p_{\text{post}})\|\mathbb{B}(p_{\text{prior}}))]}{c(u_t^i)}\right)$

Figure 2: *Overview of ACTMED.* A clinician queries the system (Step 1), prompting the agent to estimate prior disease risk. The agent simulates test outcomes (Step 2), updates beliefs (Step 3), and computes expected KL divergence to select the most informative test (Step 4). The clinician reviews and conducts the test (Steps 5–6), updates the context (Step 7), and the process repeats iteratively.

test $u_t^{(i)} \in \mathcal{U}_t$, the agent must sample plausible outcomes $u_t^{(i,j)} \sim P(u_t^{(i)} \mid K_t)$ and compute the corresponding posterior probability $P(y_{d_t} = 1 \mid K_t, u_t^{(i,j)})$ to calculate the expected information gain. Selecting appropriate samples $u_t^{(i,j)}$ from the unknown distributions and then calculating the posterior probabilities $P(y_{d_t} = 1 \mid K_t, u_t^{(i,j)})$ is non-trivial for many real-world tasks and requires accurate generative models. Physical modelling of the systems has shown good results for BED in fields like engineering [46], physics [47] and neuroscience [48]. However, in many biomedical applications the systems become too complicated to use BED with physical modelling [49].

LLM-driven sampling for ACTMED. To address sampling from complex biomedical systems, we propose using Large Language Models (LLMs) as data-driven simulators that encode rich priors and implicitly capture clinical knowledge learned from large-scale biomedical corpora. Rather than requiring explicit physical or mechanistic models, LLMs can be prompted to generate plausible samples from the joint distribution of patient variables, thereby supporting posterior inference over unobserved variables in a zero-shot setting. Prior work has shown that LLMs can anticipate clinical test outcomes and model patient trajectories with high accuracy [50], suggesting they encode useful inductive biases that can be harnessed for probabilistic reasoning in complex diagnostic tasks.

We define the utility of acquiring diagnostic feature u_t^i as the information gained per unit cost. Defining $I(u_t^i) := \mathbb{E}[\mathrm{KL}(\mathbb{B}(p_{\text{post}}) \parallel \mathbb{B}(p_{\text{prior}}))]$, we relax the dual objective (1) via a Lagrangian, assuming cost is logarithmic, i.e., $c(u_t^i) = \log c^*(u_t^i)$, as:

$$\mathcal{L}(u_t^i) = I(u_t^i) + \lambda \cdot c(u_t^i) \quad \Rightarrow \quad \mathcal{F}(u_t^i) = \frac{I(u_t^i)}{c^*(u_t^i)}. \tag{3}$$

ACTMED selects the next test to order at each time step t through the following workflow:

1. Initialize with prior belief over the diagnosis, p_{prior}, based on current knowledge K_t.
2. For each candidate diagnostic test $u_t^{(i)}$, sample M possible outcomes: $\{u_t^{(i,j)}\}_{j=1}^{M}$.
3. For each sampled outcome $u_t^{(i,j)}$, compute the corresponding posterior belief: $p_{\text{post}}^{(i,j)}$.
4. Estimate the expected KL divergence $\mathbb{E}[\mathrm{KL}(\mathbb{B}(p_{\text{post}}) \parallel \mathbb{B}(p_{\text{prior}}))]$ for each test and calculate its utility $\mathcal{F}(u_t^{(i)})$.
5. Query the test with the highest utility, $u_t^{(i^*)}$, to observe the true outcome $\hat{u}_t^{(i^*)}$.
6. Update the knowledge base with the new observation: $K_{t+1} \leftarrow K_t \cup \{\hat{u}_t^{(i^*)}\}$.

Figure 2 illustrates how ACTMED supports clinician-driven diagnostic reasoning.

Deciding when to stop information acquisition. Our KL divergence-based criterion naturally supports adaptive test acquisition by recommending a new test only when it is expected to significantly update the current belief. At each step, the agent maintains a disease belief $p_{\text{prior}} \in [0, 1]$. Diagnosis proceeds until this belief is sufficiently confident, measured relative to a decision threshold $\theta = 0.5$, regardless of the expected results of any further tests. We define the confidence gap as $\delta = |p_{\text{prior}} - \theta|$, and set a target posterior belief $q_{\text{target}} = \theta \pm \gamma\delta$, where $0 \leq \gamma \leq 1$ is a hyperparameter that controls the desired confidence margin before stopping. The sign is chosen to move the target posterior toward the decision boundary θ; the hyperparameter γ controls how much of that distance is required to justify acquiring the test. A test is acquired only if the expected KL divergence from the current belief to the candidate posterior satisfies: $\mathbb{E}[\text{KL}(\mathbb{B}(p_{\text{post}}) \parallel \mathbb{B}(p_{\text{prior}}))] \geq \mathbb{E}[\text{KL}(\mathbb{B}(q_{\text{target}}) \parallel \mathbb{B}(p_{\text{prior}}))]$. Further tests are acquired only if at least one remaining test is expected to meaningfully shift the current belief; otherwise, the model is sufficiently confident that no additional information, regardless of outcome, would alter the prediction.

Mitigating LLM hallucinations. We utilize LLMs as surrogate models for BED, employing structured prompts with three components: ① **Context specification**: A brief description of the clinical scenario and disease. ② **Known information** (K_t): Clinical observations and test results available at time t formatted in a clinical vignette (see Appendix C). ③ **Task-specific instruction**: Directives for the model's output, such as predicting test outcomes $u_t^{(i,j)} \sim P(u_t^i)$, diagnosis probabilities $P(y_{d_t} = 1 \mid K_t)$, or selecting the most informative next test. Full prompt examples are given in Appendix D.

Encouraging diverse test result sampling. We enhance model robustness and capture diagnostic uncertainty with three strategies: ① **Avoiding population averages**: The model is prompted to sample from a broader distribution of possible outcomes. ② **Increased sampling temperature**: This introduces greater randomness, reflecting higher uncertainty and improving prediction diversity. ③ **Sampling both disease presence and absence**: The model is instructed to sample outcomes under both conditions to ensure balanced and varied predictions.

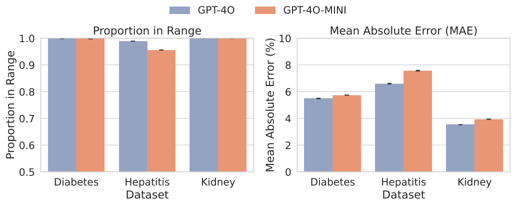

Figure 3: *Model sampling performance.* Bars show mean predictive performance over five seeds; error bars indicate standard deviation.

4 Experiments

We test two hypotheses in this study, which result from the discussions contained in the previous sections:

- **H1)** LLMs can accurately predict the distributions of diagnostic tests based on patient data.
- **H2)** LLM-based Bayesian Experimental Design (BED) improves diagnostic accuracy and efficiency through adaptive test selection.

To evaluate these hypotheses, we consider three tasks of increasing difficulty, each reflecting a clinically significant diagnostic challenge:

Chronic Kidney Disease (CKD) affects over 700 million people globally and causes more than 3 million deaths annually. Early detection is crucial for slowing disease progression and reducing mortality [51]. **Hepatitis C** infects 57 million people and leads to 300,000 deaths per year. Diagnosis is often delayed due to reliance on specialized tests, highlighting the importance of alternative screening methods like blood-based surrogates [52], [53]. **Diabetes**, affecting 500 million people worldwide, is a leading cause of cardiovascular disease and mortality. Early identification through routine health checks is critical for timely interventions [54].

4.1 LLMs accurately predict distributions of diagnostic tests

Surrogate sampling evaluation. The effectiveness of our Bayesian diagnostic framework hinges on the correctness of LLM-generated surrogate samples. We prompted models to generate $M = 10$ numerical samples per feature for each patient and compared them with real-world distributions. 98.4% of samples generated with GPT-4o-mini and 99.6% of samples generated with GPT-4o fell within empirically observed ranges, indicating effective prompt constraints against hallucination (see Fig. 3). Mean absolute percentage error (MAE) between generated values and ground truth remained below 8% between the data sets, with the more complex GPT-4o model exhibiting better performance. Feature analysis shows a high precision in general, with a slightly elevated error on high-variance biomarkers such as ALT, ALP, CHE, and serum creatinine, characteristics that are known to vary substantially between patients, especially in pathological conditions (see Fig. 4).

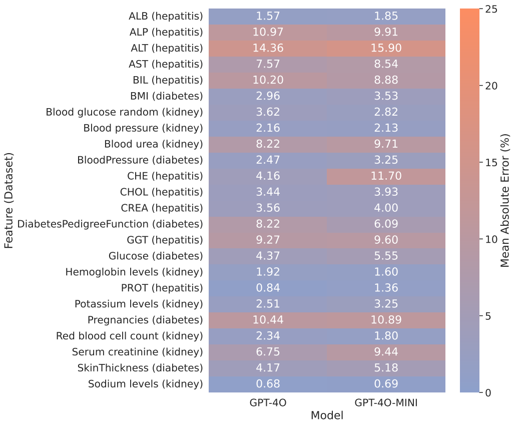

Figure 4: *Individual feature sampling performance.* Heatmap showing the best MAE percentage for each diagnostic test averaged across 5 seeds.

We emphasize that MAE is computed as the minimum distance between the true feature value and the closest generated sample, reflecting the generation of physiologically plausible values rather than precise point predictions. This evaluation aligns with the LLM's objective: to produce realistic samples within clinically feasible ranges, such that at least some closely approximate the true measurement. Performance improved with larger sample sizes (see Appendix E), suggesting benefits from ensemble generation. Full distributions in Appendix E further confirm that models generate clinically diverse samples rather than regressing to dataset means.

4.2 LLM-based BED improves diagnostic accuracy and efficiency

Timely diagnosis under resource constraints. We evaluate diagnostic accuracy under the constraint of acquiring only three clinical tests per patient, simulating real-world limitations such as acquisition delays and resource costs. Although our framework can accommodate arbitrary test cost functions, we use uniform costs across tasks to maintain consistency, as defining task-specific costs would require expert clinical input. Importantly, ACTMED is model-agnostic: its performance depends on the quality of the underlying surrogate model rather than any specific LLM architecture. We validate this by applying it across models of varying capacity; GPT-4o-mini and GPT-4o.

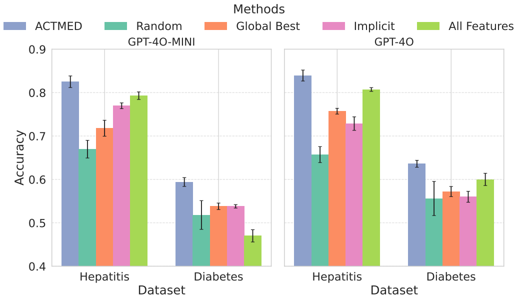

Figure 5: *Model accuracy across methods and datasets.* Bars show mean accuracy over five seeds; error bars indicate standard deviation.

We benchmark ACTMED against the following baselines that use the same models and risk prediction prompts to show how our approach can improve the performance of specific models:

- ➢ LLM classifier: No three-feature constraint; uses **all features** for direct classification.
- ➢ **Random** selection: Picks three features randomly as a stochastic baseline.
- ➢ **Global best** fixed subset: Selects a predefined set of three features prior to observing individual patient data for the task, mimicking diagnostic guidelines.
- ➢ **Implicit** selection: Selects three features actively using LLM reasoning at each step without Bayesian modelling.

Chronic Kidney Disease (CKD). Chronic kidney disease is typically diagnosed using biomarkers such as serum creatinine or glomerular filtration rate (GFR), with well-established clinical thresholds [51]. As a result, this represents a relatively straightforward classification task. Both models achieved near-perfect accuracy, even under feature selection constraints, indicating that LLMs possess strong prior knowledge of CKD's clinical presentation. This highlights their potential to support diagnostic decision-making in settings where key features are well understood. Consequently, we focus subsequent analysis on more challenging tasks, such as hepatitis and diabetes, where diagnostic ambiguity is higher and performance depends more heavily on effective test selection. Full performance metrics for all datasets and feature selection evaluation are provided in Appendix E.

Hepatitis. This task focuses on diagnosing hepatitis C using liver function tests [52]. Unlike CKD, hepatitis C often produces subtle and non-specific alterations in blood biomarkers, which can also be influenced by a range of other conditions. In our experiments, ACTMED consistently outperformed other feature selection methods across all model types (see Figure 5). Small improvements in GPT-4o compared to GPT-4o-mini are also noted. Interestingly, it even surpassed the full-information baseline, indicating that the targeted selection of informative features can improve diagnostic precision. This effect is visualized in Figure 6, which shows a clear increase in the average diagnostic accuracy across all datasets using ACTMED as additional features are acquired sequentially.

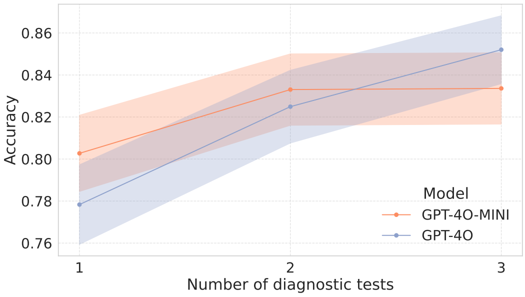

Figure 6: *Sequential diagnosis refinement.* At each step, a new test is acquired to improve accuracy, which is averaged across datasets. Errorbars denote standard deviation.

Diabetes. Diagnosing diabetes is the most challenging of the three tasks, as it lacks a single definitive test. Instead, diagnosis relies on synthesizing multiple indirect indicators such as blood glucose levels, body mass index (BMI), and blood pressure to assess overall risk [54]. Across both models, ACTMED consistently outperformed all other baselines, including the full-information setting. Performance further improved with the more capable GPT-4o model, highlighting the benefit of stronger underlying surrogate models. These results suggest that deliberate, targeted test acquisition not only enhances diagnostic accuracy but also improves generalization by reducing the influence of irrelevant or misleading features [55].

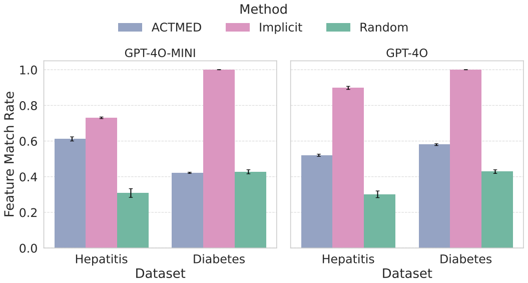

Figure 7: *Feature selection analysis.* We compare the average frequency and standard deviation of the selected features across 5 random seeds against the three globally optimal features.

4.3 Implicit Selection Lacks Personalization

For the datasets where the baseline feature selection by the model performed notably worse than ACTMED, we further analysed the tests selected by each method by evaluating how frequently models chose features identified as globally optimal prior to observing any patient-specific data (see Figure 7). Random selection serves as a stochastic baseline. Across all model-dataset pairs, implicit selection methods showed a strong bias toward globally optimal features, significantly more so than ACTMED. This effect was especially pronounced in the diabetes dataset, where both models almost exclusively selected globally optimal features, despite ACTMED achieving higher predictive accuracy. These findings suggest that LLM-based implicit selection may struggle to capture patient-specific uncertainty and adaptively personalize test acquisition, relying instead on prior knowledge about general test utility. A more detailed feature selection analysis is performed in Appendix E.

4.4 KL-Based Stopping Minimizes Redundant Testing

ACTMED not only provides greater transparency than implicit feature selection methods by generating intermediate outputs and utilizing Bayesian test selection, but also introduces a principled stopping criterion for diagnostic testing based on the expected shift in the posterior distribution, measured by KL divergence. A representative example in Table 1 shows that KL divergence consistently decreases across selection rounds, naturally allowing the stopping criterion to trigger when the expected informational gain becomes negligible. We evaluated this criterion using thresholds $\gamma \in \{0.3, 0.5, 0.7\}$, which represent varying levels of evidence required to continue testing. Baselines based on implicitly selected features or global optima did not allow early stopping and had access to all three selected tests. Our KL-based stopping method achieved superior accuracy with significantly fewer diagnostic steps (see Figure 8). As expected, higher values of γ reduce the number of diagnostic tests selected. At the conservative threshold ($\gamma = 0.6$), it reduced the average number of tests to under two across all datasets and models, except for the hepatitis dataset with GPT-4o. The method reduces overall diagnostic burden by nearly 50% while providing comparable or superior accuracy compared to baseline feature selection (see Table 2).

Table 1: Representative feature selection rounds from the Hepatitis C dataset. Darker blue indicates higher KL divergence.

Selection 1	Selection 2	Selection 3
ALB: 0.066	ALB: 0.014	ALB: 0.016
ALP: 0.041	ALP: 0.028	—
ALT: 0.172	ALT: 0.014	ALT: 0.019
AST: 0.316	—	—
BIL: 0.191	BIL: 0.025	BIL: 0.012
CHE: 0.137	CHE: 0.016	CHE: 0.015
CHOL: 0.124	CHOL: 0.020	CHOL: 0.016
CREA: 0.084	CREA: 0.021	CREA: 0.016
GGT: 0.138	GGT: 0.006	GGT: 0.018
PROT: 0.126	PROT: 0.020	PROT: 0.020

5 Discussion

BED improves LLM-based clinical diagnosis. ACTMED, integrating Bayesian Experimental Design (BED) decision-making with large language models (LLMs), outperforms naive LLMs feature selection methods in accuracy, cost-awareness, personalization, and interpretability across several datasets. In the hepatitis and diabetes datasets it even surpasses the baseline model that uses all available features. While this may seem counter-intuitive, it aligns with traditional machine learning findings where feature selection reduces noise and over-fitting [56]. Implicit sequential and global selection strategies provide strong baselines but occasionally fail to identify the most informative features. We hypothesize this arises from the model's difficulty in accurately representing how test results influence diagnostic risk due to limited domain-specific data. By shifting reasoning from the input space to the solution space, our framework enables better probabilistic inference for test selection, improving decision-making under resource constraints and leveraging LLMs' strength in contextual prediction while compensating for their limitations in Bayesian reasoning.

Comparisons to other diagnostic frameworks. Conventional clinical diagnosis relies on rule-based, population-derived guidelines that generalize well but often lack personalization and may delay detection, especially in asymptomatic patients [57]. While machine learning (ML) and deep learning (DL) approaches promise earlier detection, they are typically task-specific, lack individualized reasoning, and offer limited transparency. Even interpretability techniques such as attention maps fall short in addressing the broader need for explainable and collaborative decision-making. Recent efforts have begun applying LLMs to clinical diagnosis, but naive implementations have proven inadequate [38], particularly in our study, where LLMs struggled with personalization and had limited ability to assess the effectiveness of candidate tests. Our framework addresses these challenges by leveraging the generative strengths of LLMs within a BED framework. This enables explicit reasoning under uncertainty, personalized diagnostic pathways, and principled test selection based on expected information gain. Importantly, it also supports transparency by incorporating clinicians in the process [39]. A summary comparison of capabilities is presented in Table 3.

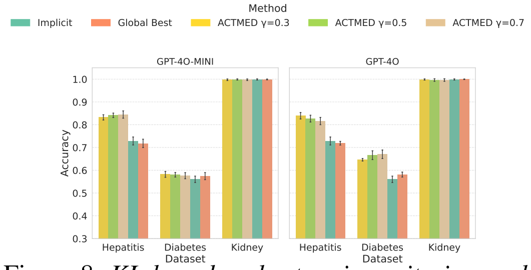

Figure 8: *KL-based early stopping criterion evaluation.* Bars represent the mean accuracy across 5 seeds, with error bars indicating standard deviation.

Table 2: Average number of tests selected under KL-based termination for varying γ values. Higher γ requires stronger evidence to continue testing. Results are reported as mean ± standard deviation across 5 random seeds.

Model	γ	Hepatitis	Diabetes	Kidney
4O-Mini	0.3	2.36 ± 0.68	2.40 ± 0.73	1.60 ± 0.75
	0.5	1.87 ± 0.77	2.06 ± 0.75	1.48 ± 0.68
	0.7	1.48 ± 0.73	1.85 ± 0.73	1.28 ± 0.52
4O	0.3	2.67 ± 0.60	2.25 ± 0.81	1.35 ± 0.64
	0.5	2.41 ± 0.67	1.90 ± 0.88	1.09 ± 0.35
	0.7	2.27 ± 0.65	1.67 ± 0.83	1.04 ± 0.24

Clinical relevance. Our framework's key advantage over standard end-to-end classifiers is its ability to involve clinicians directly in the diagnostic process. At each decision step, clinicians can review simulated test outcomes and assess their impact on diagnostic probabilities (see Appendix E). The framework issues test recommendations rather than fixed decisions, allowing clinicians to override suggestions based on context and re-query with alternative test results. While capable of providing a final disease prediction, the model serves primarily as a decision-support tool, leaving diagnostic judgment to the clinician. By highlighting the most informative tests, it streamlines clinical reasoning and reduces cognitive burden. We emphasize that current LLMs are not yet ready for direct autonomous diagnosis but may be better suited as decision-support tools, assisting clinicians in structuring and refining the diagnostic process. Importantly, safety can be enhanced by constraining the model to recommend only tests that are clinically approved for the suspected conditions, ensuring alignment with established diagnostic pathways and regulatory guidelines.

Table 3: Comparison of clinical reasoning capabilities. Only our method satisfies all criteria for transparent, timely diagnosis under resource constraints.

Capability	Rule	ML	DL	Naive LLM	**Ours**
Timely diagnosis	✗	✓	✓	✓	✓
Personalized reasoning	✗	✗	✗	✗	✓
Resource-constrained	✓	✗	✗	✗	✓
Zero-shot generalization	✓	✗	✗	✓	✓
Unstructured data	✗	✗	✗	✓	✓
Transparent explanations	✓	(✓)	✗	✗	✓

Limitations. Our framework is currently limited to binary classification. Future work should extend evaluation to multi-label datasets and co-morbidities for broader clinical applicability. The diagnostic process considers only categorical and numerical features, with free-text outputs (e.g., imaging or pathology reports) not yet incorporated, though structured versions could be included. While LLMs can interpret free-text, this adds complexity beyond our current structured approach. Additionally, our method requires more LLM queries than simpler heuristics, which may pose challenges in resource-constrained settings. However, in high-stakes domains like healthcare, the added computational cost is justified by improved decision-making, reduced uncertainty, and minimized diagnostic delays and unnecessary tests. LLMs are often criticized for their black-box nature and susceptibility to hallucinations. While these issues persist, our framework mitigates them by generating interpretable intermediate outputs, such as sampled feature distributions and uncertainty-aware predictions, providing greater transparency in the decision process. Biases in model behaviour also remain a concern, particularly in under-represented patient populations. However, by exposing the reasoning process, such as how specific test results influence diagnostic probabilities, clinicians can better identify and assess potential biases in real time, enabling more informed oversight and corrective action when needed.

Conclusions and Impact. We present ACTMED, a probabilistic framework for clinical diagnosis that uses sequential, information-theoretic decision-making to refine beliefs about disease states. Unlike traditional methods, our approach actively selects and interprets tests to inform decision-making, with LLMs acting as flexible simulators that predict test outcomes and update beliefs without task-specific training. The method is model-agnostic; its performance depends on the quality of the underlying surrogate model rather than any specific LLM architecture. While further validation is needed for clinical deployment, our system serves as a decision support tool, leaving final decisions to clinicians. As LLMs improve, frameworks like ours could enable clinician-in-the-loop systems that optimize decisions, enhance interpretability, and address resource constraints in healthcare. A promising future direction is to benchmark different LLMs to understand how architectural and training differences impact performance.

References

[1] A. Jutel, "Sociology of diagnosis: A preliminary review," *Sociology of Health &amp; Illness*, vol. 31, no. 2, pp. 278–299, Feb. 2009, ISSN: 1467-9566.

[2] Z. I. Vally, R. A. Khammissa, G. Feller, R. Ballyram, M. Beetge, and L. Feller, "Errors in clinical diagnosis: A narrative review," *Journal of International Medical Research*, vol. 51, no. 8, Aug. 2023, ISSN: 1473-2300.

[3] D. E. Newman-Toker, N. Nassery, A. C. Schaffer, *et al.*, "Burden of serious harms from diagnostic error in the usa," *BMJ Quality &amp; Safety*, vol. 33, no. 2, pp. 109–120, Jul. 2023, ISSN: 2044-5423.

[4] J. L. J. M. Müskens, R. B. Kool, S. A. van Dulmen, and G. P. Westert, "Overuse of diagnostic testing in healthcare: A systematic review," *BMJ Quality &amp; Safety*, vol. 31, no. 1, pp. 54–63, May 2021, ISSN: 2044-5423.

[5] A. U. Truth, "No health without a workforce," *World health Organisation (WHO) report*, vol. 2013, pp. 1–104, 2013.

[6] J. Hällqvist, M. Bartl, M. Dakna, *et al.*, "Plasma proteomics identify biomarkers predicting parkinson's disease up to 7 years before symptom onset," *Nature Communications*, vol. 15, no. 1, Jun. 2024, ISSN: 2041-1723.

[7] S. Fouladvand, F. R. Gomez, H. Nilforoshan, *et al.*, "Graph-based clinical recommender: Predicting specialists procedure orders using graph representation learning," *Journal of Biomedical Informatics*, vol. 143, p. 104 407, Jul. 2023, ISSN: 1532-0464.

[8] M. Khalifa and M. Albadawy, "Artificial intelligence for clinical prediction: Exploring key domains and essential functions," *Computer Methods and Programs in Biomedicine Update*, vol. 5, p. 100 148, 2024, ISSN: 2666-9900.

[9] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, *et al.*, "Revolutionizing healthcare: The role of artificial intelligence in clinical practice," *BMC Medical Education*, vol. 23, no. 1, Sep. 2023, ISSN: 1472-6920.

[10] M. Liu, S. Li, H. Yuan, *et al.*, "Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques," *Artificial Intelligence in Medicine*, vol. 142, p. 102 587, Aug. 2023, ISSN: 0933-3657.

[11] L. Tikhomirov, C. Semmler, M. McCradden, R. Searston, M. Ghassemi, and L. Oakden-Rayner, "Medical artificial intelligence for clinicians: The lost cognitive perspective," *The Lancet Digital Health*, vol. 6, no. 8, e589–e594, Aug. 2024, ISSN: 2589-7500.

[12] P. G. Shekelle, S. H. Woolf, M. Eccles, and J. Grimshaw, "Clinical guidelines: Developing guidelines," *BMJ*, vol. 318, no. 7183, pp. 593–596, Feb. 1999, ISSN: 1468-5833.

[13] S. Misra and J. H. Barth, "How good is the evidence base for test selection in clinical guidelines?" *Clinica Chimica Acta*, vol. 432, pp. 27–32, May 2014, ISSN: 0009-8981.

[14] C. Koch, K. Roberts, C. Petruccelli, and D. J. Morgan, "The frequency of unnecessary testing in hospitalized patients," *The American Journal of Medicine*, vol. 131, no. 5, pp. 500–503, May 2018, ISSN: 0002-9343.

[15] D. J. Erwin, C. LaMaire, A. Espana, T. N. Eble, and S. U. Dhar, "Financial barriers in a county genetics clinic: Problems and solutions," *Journal of Genetic Counseling*, vol. 29, no. 4, pp. 678–688, Apr. 2020, ISSN: 1573-3599.

[16] M. M. Islam, T. N. Poly, H.-C. Yang, and Y.-C. (Li, "Deep into laboratory: An artificial intelligence approach to recommend laboratory tests," *Diagnostics*, vol. 11, no. 6, p. 990, May 2021, ISSN: 2075-4418.

[17] H. Lakkaraju, D. Slack, Y. Chen, C. Tan, and S. Singh, "Rethinking explainability as a dialogue: A practitioner's perspective," *NeurIPS Workshop on Human Centered AI*,

[18] Z. Sadeghi, R. Alizadehsani, M. A. CIFCI, *et al.*, "A review of explainable artificial intelligence in healthcare," *Computers and Electrical Engineering*, vol. 118, p. 109 370, Aug. 2024, ISSN: 0045-7906.

[19] E. P. Balogh, B. T. Miller, and J. R. Ball, Eds., *Improving Diagnosis in Health Care*. National Academies Press, Dec. 2015, ISBN: 9780309377690.

[20] D. Crosby, S. Bhatia, K. M. Brindle, *et al.*, "Early detection of cancer," *Science*, vol. 375, no. 6586, Mar. 2022, ISSN: 1095-9203.

[21] A. Kumar, D. Roberts, K. E. Wood, *et al.*, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock*," *Critical Care Medicine*, vol. 34, no. 6, pp. 1589–1596, Jun. 2006, ISSN: 0090-3493.

[22] T. Schubert, R. W. Peck, A. Gimson, C. Davtyan, and M. van der Schaar, *A foundational framework and methodology for personalized early and timely diagnosis*, 2023.

[23] Y. Tang, Y. Zhang, and J. Li, "A time series driven model for early sepsis prediction based on transformer module," *BMC Medical Research Methodology*, vol. 24, no. 1, Jan. 2024, ISSN: 1471-2288.

[24] D. J. Park, M. W. Park, H. Lee, Y.-J. Kim, Y. Kim, and Y. H. Park, "Development of machine learning model for diagnostic disease prediction based on laboratory tests," *Scientific Reports*, vol. 11, no. 1, Apr. 2021, ISSN: 2045-2322.

[25] P. G. Mikhael, J. Wohlwend, A. Yala, *et al.*, "Sybil: A validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography," *Journal of Clinical Oncology*, vol. 41, no. 12, pp. 2191–2200, Apr. 2023, ISSN: 1527-7755.

[26] X. Liu, L. Faes, A. U. Kale, *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis," *The Lancet Digital Health*, vol. 1, no. 6, e271–e297, Oct. 2019, ISSN: 2589-7500.

[27] A. M. Alaa and M. van der Schaar, "Attentive state-space modeling of disease progression," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.

[28] A. M. Alaa, J. Yoon, S. Hu, and M. van der Schaar, "Personalized risk scoring for critical care prognosis using mixtures of gaussian processes," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 1, pp. 207–218, Jan. 2018, ISSN: 1558-2531.

[29] J. Jiang, W. Yang, E. M. Schnellinger, S. E. Kimmel, and W. Guo, "Dynamic logistic state space prediction model for clinical decision making," *Biometrics*, vol. 79, no. 1, pp. 73–85, Nov. 2021, ISSN: 1541-0420.

[30] A. M. Alaa and M. van der Schaar, "Balancing suspense and surprise: Timely decision making with endogenous information acquisition," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016.

[31] H. von Kleist, A. Zamanian, I. Shpitser, and N. Ahmidi, "Evaluation of active feature acquisition methods for time-varying feature settings," *Journal of Machine Learning Research*, vol. 26, no. 60, pp. 1–84, 2025.

[32] J. Clusmann, F. R. Kolbinger, H. S. Muti, *et al.*, "The future landscape of large language models in medicine," *Communications Medicine*, vol. 3, no. 1, Oct. 2023, ISSN: 2730-664X.

[33] K. Singhal, S. Azizi, T. Tu, *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Jul. 2023, ISSN: 1476-4687.

[34] D. Brin, V. Sorin, E. Konen, G. Nadkarni, B. S. Glicksberg, and E. Klang, "How gpt models perform on the united states medical licensing examination: A systematic review," *Discover Applied Sciences*, vol. 6, no. 10, Sep. 2024, ISSN: 3004-9261.

[35] H. Zong, R. Wu, J. Cha, *et al.*, "Large language models in worldwide medical exams: Platform development and comprehensive analysis," *Journal of Medical Internet Research*, vol. 26, e66114, Dec. 2024, ISSN: 1438-8871.

[36] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22, New Orleans, LA, USA: Curran Associates Inc., 2022, ISBN: 9781713871088.

[37] H. Meshkin, J. Zirkle, G. Arabidarrehdor, *et al.*, "Harnessing large language models' zero-shot and few-shot learning capabilities for regulatory research," *Briefings in Bioinformatics*, vol. 25, no. 5, Jul. 2024, ISSN: 1477-4054.

[38] P. Hager, F. Jungmann, R. Holland, *et al.*, "Evaluation and mitigation of the limitations of large language models in clinical decision-making," *Nature Medicine*, vol. 30, no. 9, pp. 2613–2622, Jul. 2024, ISSN: 1546-170X.

[39] T. Savage, A. Nayak, R. Gallo, E. Rangan, and J. H. Chen, "Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine," *npj Digital Medicine*, vol. 7, no. 1, Jan. 2024, ISSN: 2398-6352.

[40] L. Qiu, F. Sha, K. Allen, Y. Kim, T. Linzen, and S. van Steenkiste, *Bayesian teaching enables probabilistic reasoning in large language models*, 2025.

[41] X. Chen, H. Yi, M. You, *et al.*, "Enhancing diagnostic capability with multi-agents conversational large language models," *npj Digital Medicine*, vol. 8, no. 1, Mar. 2025, ISSN: 2398-6352.

[42] T. Liu, N. Astorga, N. Seedat, and M. van der Schaar, "Large language models to enhance bayesian optimization," in *The Twelfth International Conference on Learning Representations*, 2024.

[43] T. Liu, N. Huynh, and M. van der Schaar, "Decision tree induction through LLMs via semantically-aware evolution," in *The Thirteenth International Conference on Learning Representations*, 2025.

[44] N. Huynh, K. Kacprzyk, R. M. Sheridan, D. L. Bentley, and M. van der Schaar, "Decision tree induction with dynamic feature generation: A framework for interpretable DNA sequence analysis," in *Learning Meaningful Representations of Life (LMRL) Workshop at ICLR 2025*, 2025.

[45] K. Kobalczyk, N. Astorga, T. Liu, and M. van der Schaar, *Active task disambiguation with llms*, 2025.

[46] C. Papadimitriou, "Optimal sensor placement methodology for parametric identification of structural systems," *Journal of Sound and Vibration*, vol. 278, no. 4–5, pp. 923–947, Dec. 2004, ISSN: 0022-460X.

[47] S. Dushenko, K. Ambal, and R. D. McMichael, "Sequential bayesian experiment design for optically detected magnetic resonance of nitrogen-vacancy centers," *Physical Review Applied*, vol. 14, no. 5, Nov. 2020, ISSN: 2331-7019.

[48] B. Shababo, B. Paige, A. Pakman, and L. Paninski, "Bayesian inference and online experimental design for mapping neural microcircuits," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26, Curran Associates, Inc., 2013.

[49] A. White, M. Tolman, H. D. Thames, H. R. Withers, K. A. Mason, and M. K. Transtrum, "The limitations of model-based experimental design and parameter estimation in sloppy systems," *PLOS Computational Biology*, vol. 12, no. 12, A. Csikász-Nagy, Ed., e1005227, Dec. 2016, ISSN: 1553-7358.

[50] N. Makarov, M. Bordukova, R. Rodriguez-Esteban, F. Schmich, and M. P. Menden, "Large language models forecast patient health trajectories enabling digital twins," Jul. 2024.

[51] A. Francis, M. N. Harhay, A. C. M. Ong, *et al.*, "Chronic kidney disease and the global public health agenda: An international consensus," *Nature Reviews Nephrology*, vol. 20, no. 7, pp. 473–485, Apr. 2024, ISSN: 1759-507X.

[52] N. Abu-Freha, B. Mathew Jacob, A. Elhoashla, *et al.*, "Chronic hepatitis c: Diagnosis and treatment made easy," *European Journal of General Practice*, vol. 28, no. 1, pp. 102–108, May 2022, ISSN: 1751-1402.

[53] M. Martinello, S. S. Solomon, N. A. Terrault, and G. J. Dore, "Hepatitis c," *The Lancet*, vol. 402, no. 10407, pp. 1085–1096, Sep. 2023, ISSN: 0140-6736.

[54] K. L. Ong, L. K. Stafford, S. A. McLaughlin, *et al.*, "Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: A systematic analysis for the global burden of disease study 2021," *The Lancet*, vol. 402, no. 10397, pp. 203–234, Jul. 2023, ISSN: 0140-6736.

[55] G. Chatziveroglou, R. Yun, and M. Kelleher, *Exploring llm reasoning through controlled prompt variations*, 2025.

[56] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, Jun. 2022, ISSN: 2673-7647.

[57] WHO, "Guide to cancer early diagnosis," *World health Organisation (WHO) report*, 2017.

[58] H. Nori, Y. T. Lee, S. Zhang, *et al.*, *Can generalist foundation models outcompete special-purpose tuning? case study in medicine*, 2023.

[59] A. Toma, P. R. Lawler, J. Ba, R. G. Krishnan, B. B. Rubin, and B. Wang, *Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding*, 2023.

[60] J. Wu, W. Deng, X. Li, *et al.*, *Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs*, 2025.

[61] E. Goh, R. Gallo, J. Hom, *et al.*, "Large language model influence on diagnostic reasoning: A randomized clinical trial," *JAMA Network Open*, vol. 7, no. 10, e2440969, Oct. 2024, ISSN: 2574-3805.

[62] X. Zou, W. He, Y. Huang, *et al.*, "Ai-driven diagnostic assistance in medical inquiry: Reinforcement learning algorithm development and validation," *Journal of Medical Internet Research*, vol. 26, e54616, Aug. 2024, ISSN: 1438-8871.

[63] G. Mulier, S. Chevret, R. Lin, and L. Biard, "Bayesian optimal designs for multi-arm multi-stage phase ii randomized clinical trials with multiple endpoints," *Statistics in Biopharmaceutical Research*, vol. 16, no. 3, pp. 315–325, May 2024, ISSN: 1946-6315.

[64] A. Giovagnoli, "The bayesian design of adaptive clinical trials," *International Journal of Environmental Research and Public Health*, vol. 18, no. 2, p. 530, Jan. 2021, ISSN: 1660-4601.

[65] P. Nguyen, A. Rathod, D. Chapman, *et al.*, "Active semi-supervised learning via bayesian experimental design for lung cancer classification using low dose computed tomography scans," *Applied Sciences*, vol. 13, no. 6, p. 3752, Mar. 2023, ISSN: 2076-3417.

[66] A. Ananda Rao, M. Awale, and S. Davis, "Medical diagnosis reimagined as a process of bayesian reasoning and elimination," *Cureus*, Sep. 2023, ISSN: 2168-8184.

[67] J. Kornak and Y. Lu, "Bayesian decision analysis for choosing between diagnostic/prognostic prediction procedures," *Statistics and Its Interface*, vol. 4, no. 1, pp. 27–36, 2011, ISSN: 1938-7997.

[68] M. Kukar and C. Grošelj, "Machine learning in stepwise diagnostic process," in *Artificial Intelligence in Medicine*. Springer Berlin Heidelberg, 1999, pp. 315–325, ISBN: 9783540487203.

[69] O. Alagoz, T. Ayer, and F. S. Erenay, *Operations research models for cancer screening*, Jan. 2011.

[70] K. Ahuja, W. Zame, and M. van der Schaar, "Dpscreen: Dynamic personalized screening," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[71] O. Ginsburg, C.-H. Yip, A. Brooks, *et al.*, "Breast cancer early detection: A phased approach to implementation," *Cancer*, vol. 126, no. S10, pp. 2379–2393, Apr. 2020, ISSN: 1097-0142.

[72] N. Eisemann, S. Bunk, T. Mukama, *et al.*, "Nationwide real-world implementation of ai for cancer detection in population-based mammography screening," *Nature Medicine*, vol. 31, no. 3, pp. 917–924, Jan. 2025, ISSN: 1546-170X.

[73] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[74] J. Li, L. Wu, H. Dani, and H. Liu, "Unsupervised personalized feature selection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, ISSN: 2159-5399.

[75] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, Jun. 2017, pp. 1183–1192.

[76] N. Astorga, T. Liu, N. Seedat, and M. van der Schaar, "Partially observable cost-aware active-learning with large language models," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[77] D. Reker, "Practical considerations for active machine learning in drug discovery," *Drug Discovery Today: Technologies*, vol. 32–33, pp. 73–79, Dec. 2019, ISSN: 1740-6749.

[78] J. Piskorz, N. Astorga, J. Berrevoets, and M. van der Schaar, "Active feature acquisition for personalised treatment assignment," in *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

[79] L. Rubini, P. Soundarapandian, and P. Eswaran, *Chronic Kidney Disease*, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5G020, 2015.

[80] R. Lichtinghagen, F. Klawonn, and G. Hoffmann, *HCV data*, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5D612, 2020.

[81] J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, "Using the adap learning algorithm to forcast the onset of diabetes mellitus," *Proceedings - Annual Symposium on Computer Applications in Medical Care*, vol. 10, Nov. 1988.

[82] S. Im, J. Oh, and E. Choi, *Labtop: A unified model for lab test outcome prediction on electronic health records*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We introduce and formalize a framework for timely diagnosis using BED and LLMs in Sections 2 & 3. We validate the performance of our model on real-world medical datasets in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of the model are discussed in Section 5. Computational cost is discussed in Appendix F.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The assumptions underlying the Bayesian Experimental Design are discussed in Section 2. The paper does not include novel theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provided detailed information on the publicly available datasets as well as the prompts, pseudocode for the KL divergence based test selection and specific models used to run the experiment. All source code and reproduction instructions will be made available upon publication.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The datasets are all open access, with details provided in Section C. The experiments and prompts used are also described. All source code and reproduction instructions will be made available upon publication.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details on the datasets used and the LLM setup are provided in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We performed all experiment across 5 different random seeds and report all results as the mean of the runs with the corresponding standard deviation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The bulk on the experiments other than the API calls was performed on the Azure Open AI services, as detailed in Appendix C & F.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss how our approach may lead to more rational test selection through personalization in Section 5. We also acknowledge limitations of directly deploying LLMs in clinical practice and highlight that our model maintains a human-in-the-loop approach and clinicians can override test suggestion and review the models outputs at each step.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release any pretrained language models or datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators of the datasets as well as the licences under which these are available are provided in Appendix C.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not create novel datasets or pretrained models. All source code and reproduction instructions will be made available upon publication.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The use of LLMs as generative models for Bayesian Experimental Design is described in Section 3.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Extended related works

LLMs for clinical diagnosis

There has been growing interest in validating large language models (LLMs) on clinical diagnosis tasks, alongside efforts to fine-tune these models for biomedical datasets. Interestingly, recent work has shown that general-purpose LLMs can sometimes outperform specialized models even on domain-specific benchmarks [58]. To improve clinical reasoning, some approaches have involved training LLMs on medical dialogues [59], while others have incorporated structured knowledge sources such as medical knowledge graphs, particularly for tasks resembling medical licensing exams [60]. Beyond static evaluation, researchers have explored interactive diagnostic frameworks that allow LLMs to collaborate with clinicians. For instance, agentic networks of LLMs have been shown to enhance diagnostic reasoning compared to individual models, though their effectiveness strongly depends on the quality of the underlying language model [41]. Conversely, simply providing clinicians with access to unrefined LLMs has not yielded significant improvements in diagnostic performance [61]. Reinforcement learning has also been proposed as a way to teach models clinical reasoning behaviours, but such methods typically require task-specific training and substantial supervision [62].

Bayesian methods in medicine

Bayesian methods have long been influential in clinical research, particularly for optimizing trial design, such as adaptive patient allocation and dose-finding strategies in early-phase drug development [63], [64]. In medical imaging, Bayesian active learning approaches have been used to strategically acquire informative data points, improving model efficiency and performance [65]. More broadly, the diagnostic process itself is naturally aligned with Bayesian reasoning, where clinical evidence incrementally updates probabilistic beliefs about possible conditions [66]. In diagnostic settings, Bayesian approaches have been specifically applied to optimize test selection by leveraging the known sensitivity and specificity of various diagnostic tools [67]. Early foundational work in this domain illustrated the potential of stepwise Bayesian integration of test results, where each successive diagnostic query is strategically chosen based on its expected diagnostic value [68]. These methods effectively utilize probabilistic reasoning over potential test outcomes, exploiting the inherent accuracy profiles of available tests to make efficient decisions. Furthermore, prior work on the development of timely diagnosis tools has shown significant benefits in applying structured approaches to population-based screening efforts [69], [70]. Such diagnosis of asymptomatic patients is critical as it enables earlier intervention, which has been consistently shown to improve treatment outcomes for various diseases [22], [71]. Machine learning approaches have further enhanced the effectiveness of this screening in specific diseases, including breast and lung cancer [25], [72].

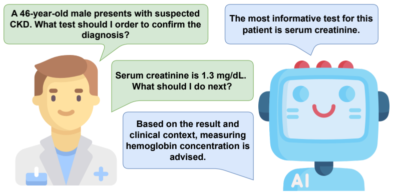
A 46-year-old male presents with suspected CKD. What test should I order to confirm the diagnosis?
The most informative test for this patient is serum creatinine.
Serum creatinine is 1.3 mg/dL. What should I do next?
Based on the result and clinical context, measuring hemoglobin concentration is advised.
AI

Figure 9: Example of a ML assistant guiding a clinician through sequential diagnostic refinement for a suspected chronic kidney disease (CKD) case. The agent recommends the most informative next test and iteratively updates its suggestions as new results arrive.

Despite these valuable applications and advances, a unified Bayesian framework that fully supports the entire sequential decision-making process inherent in clinical diagnosis, from the initial formation of hypotheses to the targeted acquisition and sophisticated integration of heterogeneous clinical data, remains largely underdeveloped. Our current work directly addresses this critical gap by demonstrating how large language models (LLMs) can effectively serve as powerful surrogates for Bayesian Experimental Design, thereby providing a novel computational approach to support and improve real-world diagnostic tasks by facilitating this comprehensive sequential reasoning and data integration process.

Active learning in medicine

Active learning is a subfield of machine learning in which models strategically select the most informative data points to observe, thereby minimizing the need for costly human labelling [73]. It has been successfully applied in both unsupervised settings to identify informative features [74] and in conjunction with deep learning architectures to improve data efficiency [75]. Central to active learning is the definition of acquisition functions or utility metrics that quantify the expected benefit of observing new data. Large language models (LLMs) have recently been explored as surrogate models for guiding such selection, particularly in settings where data acquisition is expensive or sparse [76]. In biomedicine, active learning has been applied to optimize test ordering and experimental design, where the cost of procedures is often a limiting factor. For example, in pharmacology, it has been used to prioritize compound testing [77], and in clinical treatment planning, active feature selection methods have enabled more efficient personalization of care [78]. Despite these advances, the use of LLMs for active test selection in the context of sequential clinical diagnosis remains unexplored. Figure 9 illustrates how such an ML-based framework can support clinicians by optimizing the selection of diagnostic tests to sequentially refine the diagnosis.

B Extended works

Summary of formalism

A summary of the mathematical formalism introduced is given in Table 4.

Table 4: Summary of notation used in the agent-based diagnosis and Bayesian experimental design framework.

Symbol	**Definition**
Σ	Space of natural language strings
$\mathcal{S} \in \Sigma$	Natural language input to the agent
$\mathcal{D}_t \subset \Sigma$	Set of all possible diagnoses at time t
$\mathcal{D}_{\text{true},t} \subset \mathcal{D}_t$	Subset of true diagnoses for a patient at time t
$d_t \in \mathcal{D}_t$	A single diagnosis at time t
$T = \{1, 2, \ldots, T_{\max}\}$	Discrete time horizon
$\mathcal{X}_t \subset \Sigma$	Ground truth information at time t
$K_t \subset \mathcal{X}_t$	Known (observed) information at time t
$\mathcal{U}_t = \mathcal{X}_t \setminus K_t$	Unknown information at time t
$P(y_{d_t} = 1 \mid K_t)$	Posterior belief over diagnosis d_t given knowledge at time t
$\mathbb{H}[P(\cdot)]$	Shannon entropy of the probability distribution
$u_t^{(i)}$	Candidate unobserved variable (test) at time t
$u_t^{(i,j)} \sim P(u_t^{(i)})$	Sampled outcome j from candidate test i
$\hat{u}_t^{(i)}$	Ground truth result of a variable (test) at time t
$\text{EIG}(u_t^{(i)})$	Expected information gain of test i
p_{prior}	Prior belief distribution before test observation
p_{post}	Posterior distribution after observing test result
$\text{KL}(q \parallel p)$	KL divergence between two distributions
M	Number of Monte Carlo samples
$\mathcal{F}(u_t^{(i)})$	Utility of diagnostic test i
$c(u_t^{(i)})$	Cost of test $u_t^{(i)}$

Directly calculating the EIG from entropy

The information gain (IG) from an additional piece of information $u_t^{(i,j)}$ is defined as:

$$\text{IG}(u_t^{(i)}) := \mathbb{H}[P(y_d = 1 \mid K_t)] - \mathbb{H}[P(y_d = 1 \mid K_t, u_t^{(i,j)})]. \tag{4}$$

where $\mathbb{H}$ denotes Shannon entropy. The first term represents the uncertainty about the diagnosis D_t given the current knowledge K_t, while the second term represents the uncertainty after observing the additional piece of information $u_t^{(i)}$. Since this quantity is difficult to compute directly, we approximate it using an expectation over samples to obtain the Expected Information Gain (EIG):

$$\text{EIG}(u_t^{(i)}) = \mathbb{H}[P(y_d = 1 \mid K_t)] - \mathbb{E}_{u_t^{(i,j)} \sim P(u_t^i)} \left[\mathbb{H}[P(y_d = 1 \mid K_t, u_t^{(i,j)})] \right]. \tag{5}$$

While it is possible to directly estimate the expected information gain (EIG) using changes in entropy, this approach can sometimes favour diagnostic tests that are not truly informative. For binary classification tasks, we model the posterior distribution $P(y_d = 1 \mid K_t, u_t^{(i)})$ as a Bernoulli distribution $\mathbb{B}(p_i)$ with success probability $p_i \in [0, 1]$. The entropy of a Bernoulli distribution is given by:

$$\mathbb{H}[\mathbb{B}(p_i)] = -p_i \log(p_i) - (1 - p_i) \log(1 - p_i). \tag{6}$$

We approximate the expected entropy using samples $u_t^{(i,j)} \sim P(\mathcal{U}_t)$ and the corresponding posterior probabilities $p_{i,j} = P(y_d = 1 \mid K_t, u_t^{(i,j)})$:

$$\text{EIG}(u_t^{(i)}) \approx \mathbb{H}[P(y_d = 1 \mid K_t)] - \frac{1}{M} \sum_{j=1}^{M} \left(-p_{i,j} \log p_{i,j} - (1 - p_{i,j}) \log(1 - p_{i,j}) \right). \tag{7}$$

Finally, the optimal piece of information to query is:

$$i^* = \arg\max_i \text{EIG}(u_t^{(i)}). \tag{8}$$

However, this entropy-based formulation may favour tests that reinforce already confident but potentially incorrect predictions, rather than tests that challenge or refine them. For instance, consider a patient for whom the model initially assigns a low disease probability of 0.1. Suppose a highly informative test exists which, if performed, would increase the disease probability to 0.6. In contrast, another test may have no diagnostic value, leaving the probability unchanged at 0.1 regardless of the outcome. The entropy-based EIG could erroneously prefer the non-informative test, as the posterior remains confidently near an extreme. In contrast, a formulation based on expected Kullback–Leibler (KL) divergence captures the magnitude of change in the posterior distribution, and would correctly favor the more informative test, as it induces a larger shift in the model's belief.

C Experimental details

Datasets

We evaluated the performance of our model on three clinical datasets of varying complexity:

Chronic Kidney Disease. The first task involves the prediction of chronic kidney disease (CKD) from symptoms and laboratory results [79]. The model is provided with demographic variables such as age and gender, as well as signs observable on physical examination (e.g., oedema). It is then tasked with diagnosing CKD by selectively ordering lab tests such as serum albumin or serum creatinine. We filtered the dataset to retain only instances without missing values, resulting in a total of 157 patients. Categorical lab tests lacking a clear clinical interpretation, such as Pus Cells: Abnormal, were excluded. The dataset is available from the UCI Machine Learning Repository under a CC BY 4.0 licence: `https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease`.

Hepatitis. The second task requires the model to diagnose Hepatitis C virus (HCV) infection using liver function test results [80]. Age and sex were provided as known demographic features, while laboratory tests (e.g., ALT, AST, GGT) could be queried by the model. We included all 56 patients with confirmed HCV and selected a random subset of 56 healthy individuals to form a balanced dataset. The dataset is publicly available from the UCI Repository under a CC BY 4.0 licence: `https://archive.ics.uci.edu/dataset/571/hcv+data`.

Diabetes. The third task uses a random subset of 100 patients from the Pima Indians Diabetes dataset, originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases [81]. The model is asked to predict the presence of diabetes based on clinical measurements. All individuals are female, at least 21 years old, and of Pima Indian heritage. Age is treated as a known feature; the remaining features are unknown and can be selectively queried. The dataset is available on Kaggle under a CC0 Public Domain licence: `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`.

Input data formatting

LLMs have been shown to struggle interpreting tabular clinical data accurately [38]. To mitigate this limitation, we converted all available clinical information from structured tabular form into concise natural language vignettes. Each vignette integrates demographic information, diagnostic test results, and measurement units where appropriate, providing a more interpretable and context-rich format for the model. Table 5 presents an example of the raw tabular input used for diagnosing chronic kidney disease. This input is automatically converted into a clinical vignette using a rule-based preprocessing script (see Clinical Vignette 1).

Table 5: Structured clinical data for an example patient prior to natural language transformation and inclusion of appropriate units.

Category	**Test Name**	**Result**
Demographics	Age	63
Vital Signs	Blood Pressure	70
Urine Tests	Specific Gravity	1.010
	Albumin	3
	Sugar	0
	Red Blood Cells	Abnormal
	Pus Cells	Abnormal
	Clumps	Present
	Bacteria	Not Present
Blood Tests	Blood Glucose	380
	Blood Urea	60
	Creatinine	2.7
	Hemoglobin	10.8
	PCV	32
	WBC Count	4500
	RBC Count	3.8
Electrolytes	Sodium	131
	Potassium	4.2
Symptoms	Appetite	Poor
	Pedal Edema	Yes
	Anemia	No
Comorbidities	Hypertension	Yes
	Diabetes Mellitus	Yes
	Coronary Artery Disease	No

Clinical Vignette 1

The patient is 63 years old. The patient's diastolic blood pressure is 70 mm/Hg. The patient has a poor appetite. The patient has pedal oedema. The patient has hypertension. The patient has diabetes mellitus. The patient does not have coronary artery disease. The patient does not have anaemia. Specific gravity was measured at 1.01. Albumin levels in urine was measured at 3.0. Sugar levels in urine was measured at 0.0. Blood glucose random was measured at 380.0 mg/dL. Blood urea was measured at 60.0 mg/dL. Serum creatinine was measured at 2.7 mg/dL. Sodium levels was measured at 131.0 mEq/L. Potassium levels was measured at 4.2 mEq/L. Haemoglobin levels was measured at 10.8 g/dL. Packed cell volume was measured at 32.0.

Experiment

For each patient in a preprocessed dataset subset, we evaluated risk predictions using all features, as well as under four feature selection strategies: Bayesian, Random, Global Best (predefined), and Implicit. At each iteration, one additional feature from the unknown set was revealed, and corresponding risks were computed. Pseudocode for the Bayesian selection using the KL-divergence is given in Algorithm 1. Importantly, the risk prediction prompt remained unchanged between feature selection methods other than the clinical information added at the end. All experiments were implemented using GPT-4o (Version 2024-11-20) and GPT-4o-mini (Version 2024-07-18) as provided on the Azure OpenAI Service. To ensure robustness, each experiment was run across 5 different random seeds. For all experiments, we set the number of sampled test outcomes or risk probability distributions to 10. To ensure the sampling produced a more diverse set of responses, we used a temperature of 1 and specifically instructed the model in the prompts to simulate randomness. Performance metrics were averaged across runs with their corresponding standard deviation.

Algorithm 1 KL-guided Diagnostic Test Selection

```
Require: Initial knowledge $K_t$, unknowns $\{u_t^{(i)}\}_{i=1}^N$, number of samples $M$
Sample $M$ draws from $P(y_{d_t} = 1 \mid K_t)$: $\{p_{\text{prior}}^{(j)}\}_{j=1}^M$
for $i = 1$ to $N$ do
    for $j = 1$ to $M$ do
        Sample $u_t^{(i,j)} \sim P(u_t^{(i)})$
        Compute posterior: $p_{\text{posterior}}^{(j)} = P(y_{d_t} = 1 \mid K_t, u_t^{(i,j)})$
        Compute $\text{KL}^{(j)} = p_{\text{prior}}^{(j)} \log \frac{p_{\text{prior}}^{(j)}}{p_{\text{posterior}}^{(j)}} + (1 - p_{\text{prior}}^{(j)}) \log \frac{1-p_{\text{prior}}^{(j)}}{1-p_{\text{posterior}}^{(j)}}$
    end for
    $\mathbb{E}[\text{KL}(u_t^{(i)})] = \frac{1}{M} \sum_{j=1}^M \text{KL}^{(j)}$
end for
$i^* = \arg\max_i \mathbb{E}[\text{KL}(u_t^{(i)})]$
Request $\hat{u}_t^{(i^*)}$ and update $K_{t+1} = K_t \cup \{\hat{u}_t^{(i^*)}\}$
```

D LLM Prompts

For each dataset, we use four distinct prompt types. The first prompt, shared across all methods, is used for risk prediction based on known patient data, ensuring consistency in evaluation. The second prompt is issued at the start of the experiment, before observing any patient-specific information, to identify globally optimal features. The third prompt is used in the implicit selection baseline, where the model is queried at each decision step to choose the most suitable test based on the information observed so far. The final prompt is specific to the Bayesian Experimental Design setting and is used to sample a plausible outcome for a candidate diagnostic test, conditioned on the currently available patient data.

Risk Prediction Prompt: CKD

You are an expert nephrologist. Based on the following clinical data and the patient's history, provide an estimate of the patient having chronic kidney disease as a single number between 0 and 1. Consider key laboratory markers and other pertinent values. When these values indicate kidney disease or damage—assign a number closer to 1, indicating a higher probability of chronic kidney disease. If the laboratory values are within normal ranges, assign a value closer to 0. Return only the number that can be converted to a Python `float`, without any additional commentary.
`$known_info`

Global Feature Selection Prompt: CKD

You are an expert nephrologist. Based on the following list of features: `$all_features`, please indicate which `$n` features you believe are the most informative and critical for diagnosing chronic kidney disease.
Return your answer as a Python list of exactly `$n` feature names.
For example: if `n` is 1, return `['age']`; if `n` is 2, return `['age', 'bp']`.
Do not include any additional commentary.

Feature Selection Prompt: CKD (Implicit)

You are an expert nephrologist. Based solely on the following known clinical data, determine which additional feature from the list below would be the most informative to sample next for diagnosing chronic kidney disease.
Known Data: `$known_data`
Unknown Features: `$unknown_features`
Return only the name of the feature strictly in the form shown in the list as a string, without any additional commentary.

Sampling Prompt: CKD Lab Test Simulation

You are an expert nephrologist. Based on the following clinical data and the patient's history, please simulate a random draw from the full range of clinically plausible values for `$feature_to_sample`.
The value should not simply be the average or a central tendency, but should vary as if sampled at random from a realistic distribution. Consider the following description: `$ref_info`.
Avoid returning the same value repeatedly across multiple draws, and ensure the value varies as if sampled from a plausible distribution. Introduce randomness by considering edge cases, typical values, and outliers within the plausible range.
Return your answer as a single numeric value that can be parsed into a `float` in Python, with no additional commentary or units.
IMPORTANT: Assume that the patient may or may not have chronic kidney disease, and your sampling should reflect that uncertainty.
`$known_info`
IMPORTANT: Under NO circumstances provide explanations, commentary, or text beyond the single numeric float or string requested. The response MUST be parseable strictly as a float, e.g., 0.512, with no extra words. If a string is requested no float is required.

Risk Prediction Prompt: Hepatitis C

You are an expert hepatologist. Based on the following clinical data and the patient's history, please provide an estimate of the patient's risk of being infected with hepatitis C as a single number between 0 and 1. Consider key laboratory markers and other pertinent values. When these values indicate liver inflammation or damage — assign a number closer to 1, indicating a higher probability of hepatitis C infection. If the laboratory values are within normal ranges, assign a value closer to 0. Return only the number that can be converted to a Python float, without any additional commentary.
`$known_info`

Global Feature Importance Prompt: Hepatitis C

You are an expert hepatologist. Based on the following list of features: `$all_features`, please indicate which `$n` features you believe are the most informative and critical for diagnosing hepatitis.
Return your answer as a Python list of exactly `$n` feature names (for example, if `n` is 1, return `['ALT']`; if `n` is 2, return `['ALT', 'AST']`), without any additional commentary.
`$known_info`

Feature Selection Prompt: Hepatitis C (Implicit)

You are an expert hepatologist. Based solely on the following known clinical data, determine which additional feature from the list below would be the most informative to sample next for diagnosing hepatitis.
Known Data: `$known_data`
Unknown Features: `$unknown_features`
Return only the name of the feature as a string, without any additional commentary.

Sampling Prompt: Hepatitis C Lab Test Simulation

You are an expert hepatologist. Based on the following clinical data and the patient's history, please simulate a random draw from the full range of clinically plausible values for `$feature_to_sample`.
Consider the possible range as described: `$ref_info`. Ensure that the value you return is realistic and reflects clinical variability. Avoid returning the same value repeatedly across multiple draws, and ensure the value varies as if sampled from a plausible distribution. Introduce randomness by considering edge cases, typical values, and outliers within the plausible range.
Return your answer as a single numeric value that can be converted to a Python float, without any additional commentary.
IMPORTANT: Assume that the patient may or may not have hepatitis C, and your sampling should reflect that uncertainty.
`$known_info`
IMPORTANT: Under NO circumstances provide explanations, commentary, or text beyond the single numeric float or string requested. The response MUST be parseable strictly as a float, e.g., 0.512, with no extra words. If a string is requested no float is required.

Risk Prediction Prompt: Diabetes

You are an expert endocrinologist. Based on the following clinical data and the patient's history, provide an estimate of the patient's risk of diabetes as a single number between 0 and 1.
It is known that all patients are females at least 21 years old of Pima Indian heritage. Focus on key markers. Assign a value closer to 1 if the data indicate high risk, and closer to 0 if within normal limits. Return only the number, without any additional commentary.
`$known_info`

Global Feature Importance Prompt: Diabetes

You are an expert endocrinologist. Based on the following list of features: `$all_features`, please indicate which `$n` features you believe are the most informative and critical for diagnosing diabetes.
Return your answer as a Python list of exactly `$n` feature names (for example, if `n` is 1, return `['Glucose']`; if `n` is 2, return `['Glucose', 'BMI']`), without any additional commentary.

Feature Selection Prompt: Diabetes (Implicit)

You are an expert endocrinologist. Based solely on the following known clinical data, determine which additional feature from the list below would be the most informative to sample next for diagnosing diabetes.
Known Data: `$known_data`
Unknown Features: `$unknown_features`
Return only the name of the feature as a string, without any additional commentary.
`$known_info`

Sampling Prompt: Diabetes Lab Test Simulation

You are an expert endocrinologist. Based on the following clinical data and the patient's history, please simulate a random draw from the full range of clinically plausible values for `$feature_to_sample`.
Consider the following unit for the sampled value: `$ref_info`. Ensure that the value you return is realistic and reflects clinical variability. Avoid returning the same value repeatedly across multiple draws, and ensure the value varies as if sampled from a plausible distribution. Introduce randomness by considering edge cases, typical values, and outliers within the plausible range.
Return your answer as a single numeric value that can be converted to a Python float with no units or additional commentary.
IMPORTANT: Assume that the patient may or may not have diabetes, and your sampling should reflect that uncertainty.
`$known_info`
IMPORTANT: Under NO circumstances provide explanations, commentary, or text beyond the single numeric float or string requested. The response MUST be parseable strictly as a float, e.g., 0.512, with no extra words. If a string is requested no float is required.

E Extended results

Predicting plausible test outcome distributions

To evaluate model performance, we analyzed how the mean absolute error (MAE) changed with increasing numbers of generated samples. For computational efficiency, we limited the number of samples to 10 per query. As shown in Figure 10, we observe a consistent decrease in MAE as more features are acquired, indicating that the model is producing a diverse set of samples, some of which closely approximate the ground truth. Furthermore, the larger GPT-4o model outperformed the smaller GPT-4o-mini variant across all three datasets, highlighting the benefits of scale in both predictive accuracy and sample quality.

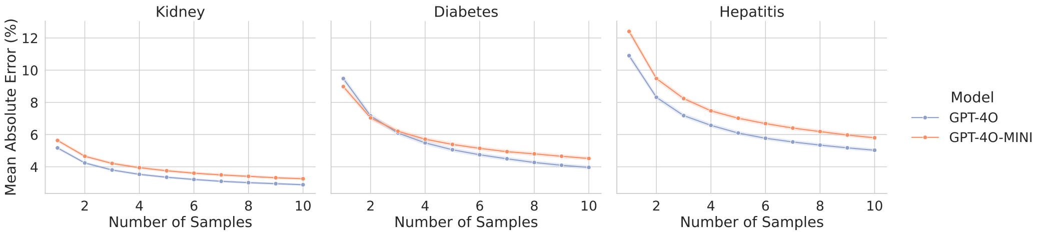

Figure 10: *Increasing sample size improves the MAE.* Model performances are averaged across 5 different seeds and are shown with their corresponding standard deviations.

To assess the ability of large language models (LLMs) to generate diverse samples from hypothetical test outcome distributions, we analysed the generated values for all numerical features across the three datasets. Categorical features were excluded from this analysis. For numerical features, the LLM produced non-deterministic outputs that varied meaningfully between samples, rather than issuing static or averaged responses. This behaviour indicates that the model is capable of representing distributional uncertainty in a clinically meaningful way. This behaviour was consistent across all three datasets, as illustrated in Figures 11, 12 and 13.

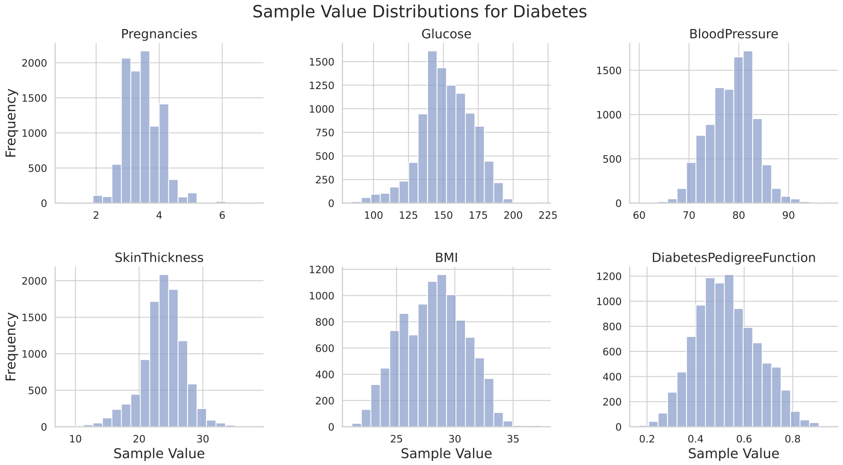

Figure 11: *Sample variability of LLM predictions in the diabetes dataset.* Each bar chart displays the distribution of values sampled for numerical features. The heterogeneity across samples indicates the model's capacity to avoid mode collapse and to reflect uncertainty consistent with clinical variation.

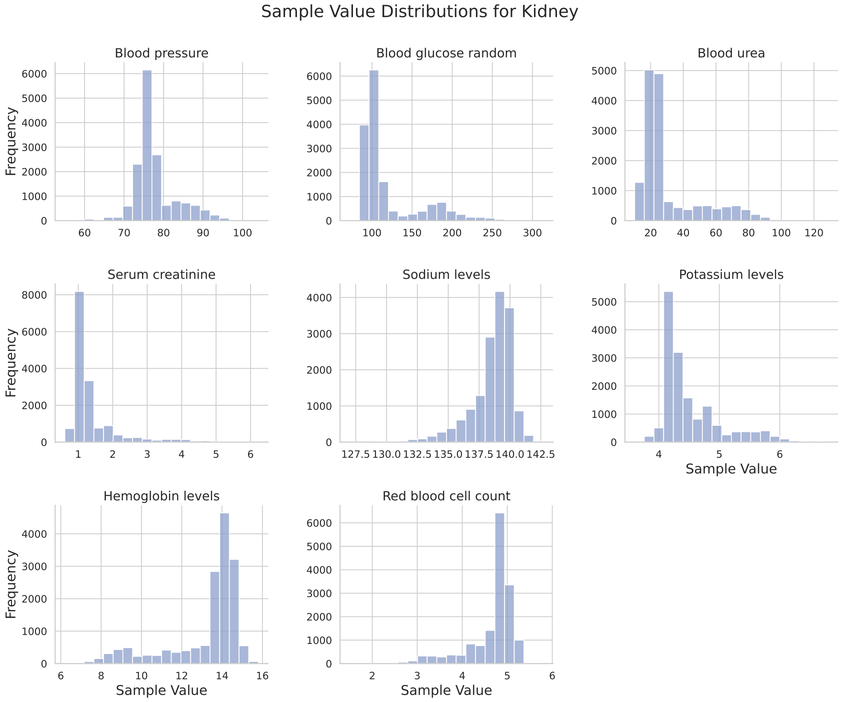

Figure 12: *Sampled feature distributions in the chronic kidney disease (CKD) dataset.* The LLM produces diverse and physiologically grounded value ranges across all features. This supports its utility as a surrogate model in capturing uncertainty for Bayesian evidence diagnostics.

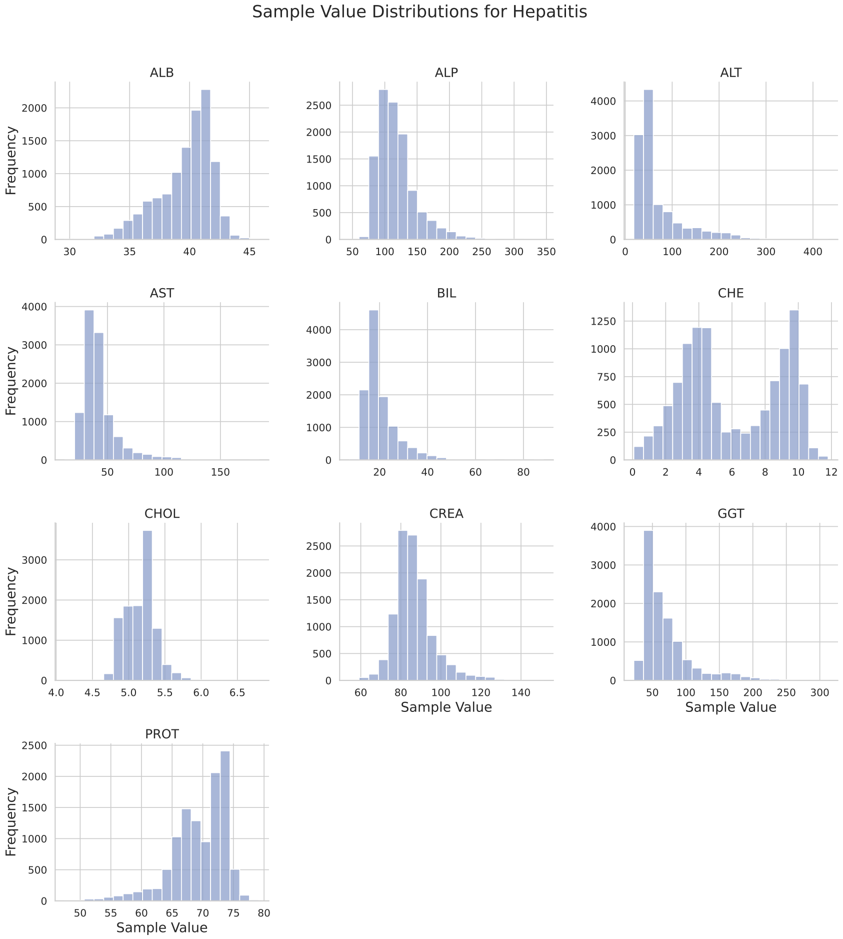

Figure 13: *Distributions of LLM-generated samples for numerical features in the hepatitis dataset.* Bar plots show the spread of sampled values for each feature, illustrating the model's ability to represent plausible clinical variability. The distributions vary across features, reflecting both physiological ranges and disease-specific uncertainty.

Diagnostic Performance Evaluation

We evaluate ACTMED against multiple baselines using accuracy, precision, recall, F1 score, and ROC-AUC (see Tables 6 and 7). Performance improves across all metrics when moving from GPT-4o-mini to the more capable GPT-4o. On the kidney dataset, differences between models are minimal, suggesting both models perform reliably. In contrast, on the hepatitis and diabetes datasets, ACTMED consistently outperforms baseline feature selection methods and the full-feature classifier, demonstrating its ability to adaptively select informative subsets.

We further assess ACTMED under varying feature budget constraints (Table 8), controlled via the stopping threshold γ. Across all datasets and both models, diagnostic accuracy remains stable even as γ varies. However, stricter thresholds (higher γ) lead to more conservative test acquisition and substantially fewer tests. For example, on the kidney dataset, both models typically require only 1–2 tests per patient when using the stopping criterion. At a conservative threshold of $\gamma = 0.7$, GPT-4o achieves near-perfect accuracy while querying just one test in nearly all cases.

Table 6: Performance metrics (mean ± std) for GPT-4o-mini across datasets. Best-performing methods are in **bold**; if All Features is best, the second-best is also indicated.

Dataset	Method	AUC	Accuracy	F1	Precision	Recall
Kidney	All Features	**1.000 ± 0.000**	0.975 ± 0.006	0.956 ± 0.009	0.915 ± 0.017	**1.000 ± 0.000**
	ACTMED	**1.000 ± 0.000**	0.992 ± 0.006	0.986 ± 0.011	0.978 ± 0.020	0.995 ± 0.009
	Random	**1.000 ± 0.000**	0.981 ± 0.004	0.966 ± 0.007	0.947 ± 0.021	0.986 ± 0.011
	Global Best	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	Implicit	**1.000 ± 0.000**	0.997 ± 0.005	0.995 ± 0.009	0.991 ± 0.018	**1.000 ± 0.000**
Hepatitis	All Features	0.874 ± 0.002	0.793 ± 0.009	0.758 ± 0.011	**0.910 ± 0.012**	0.650 ± 0.014
	ACTMED	**0.891 ± 0.006**	**0.825 ± 0.013**	**0.806 ± 0.015**	**0.903 ± 0.017**	**0.729 ± 0.017**
	Random	0.743 ± 0.018	0.670 ± 0.020	0.558 ± 0.031	0.843 ± 0.045	0.418 ± 0.029
	Global Best	0.824 ± 0.007	0.718 ± 0.018	0.668 ± 0.022	0.812 ± 0.031	0.568 ± 0.021
	Implicit	0.856 ± 0.016	0.770 ± 0.007	0.743 ± 0.005	0.843 ± 0.023	0.664 ± 0.013
Diabetes	All Features	0.732 ± 0.037	0.470 ± 0.014	0.583 ± 0.006	0.411 ± 0.006	**1.000 ± 0.000**
	ACTMED	0.759 ± 0.025	**0.594 ± 0.010**	**0.633 ± 0.010**	**0.476 ± 0.007**	0.946 ± 0.017
	Random	0.648 ± 0.057	0.518 ± 0.033	0.590 ± 0.017	0.431 ± 0.018	0.935 ± 0.022
	Global Best	**0.766 ± 0.022**	0.538 ± 0.007	0.616 ± 0.004	0.445 ± 0.004	**1.000 ± 0.000**
	Implicit	0.758 ± 0.024	0.538 ± 0.004	0.616 ± 0.002	0.445 ± 0.002	**1.000 ± 0.000**

Table 7: Performance metrics (mean ± std) for GPT-4o across datasets. Best-performing methods are in **bold**; if **All Features** is best, the second-best method is also highlighted.

Dataset	Method	AUC	Accuracy	F1	Precision	Recall
Kidney	All Features	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	Bayesian	**1.000 ± 0.000**	**0.999 ± 0.003**	**0.998 ± 0.005**	**1.000 ± 0.000**	**0.995 ± 0.009**
	Random	0.999 ± 0.001	0.995 ± 0.003	0.991 ± 0.005	**1.000 ± 0.000**	0.981 ± 0.009
	Global Best	**1.000 ± 0.000**	**0.999 ± 0.003**	**0.998 ± 0.005**	**1.000 ± 0.000**	**0.995 ± 0.009**
	Implicit	**1.000 ± 0.000**	**0.999 ± 0.003**	**0.998 ± 0.005**	**1.000 ± 0.000**	**0.995 ± 0.009**
Hepatitis	All Features	0.876 ± 0.014	**0.807 ± 0.004**	**0.771 ± 0.006**	**0.948 ± 0.001**	**0.650 ± 0.009**
	Bayesian	**0.917 ± 0.013**	**0.839 ± 0.013**	**0.814 ± 0.016**	**0.966 ± 0.012**	**0.704 ± 0.021**
	Random	0.728 ± 0.039	0.657 ± 0.018	0.507 ± 0.036	0.904 ± 0.052	0.354 ± 0.035
	Global Best	0.799 ± 0.013	0.757 ± 0.007	0.699 ± 0.009	0.919 ± 0.011	0.564 ± 0.009
	Implicit	0.805 ± 0.013	0.729 ± 0.016	0.646 ± 0.030	0.927 ± 0.008	0.496 ± 0.036
Diabetes	All Features	0.766 ± 0.014	0.600 ± 0.014	**0.640 ± 0.011**	0.480 ± 0.009	**0.962 ± 0.013**
	ACTMED	0.764 ± 0.012	**0.636 ± 0.008**	**0.636 ± 0.010**	**0.505 ± 0.007**	0.859 ± 0.046
	Random	0.664 ± 0.039	0.556 ± 0.039	0.592 ± 0.032	0.449 ± 0.026	0.870 ± 0.050
	Global Best	**0.792 ± 0.012**	0.572 ± 0.012	0.623 ± 0.006	0.462 ± 0.007	0.957 ± 0.013
	Implicit	0.773 ± 0.015	0.560 ± 0.013	0.618 ± 0.009	0.455 ± 0.008	**0.962 ± 0.013**

Table 8: Average number of tests selected and accuracy (mean ± standard deviation) under KL-based termination for varying γ values. Lower γ requires stronger evidence to continue testing. Results are reported across 5 random seeds.

Model	γ	Hepatitis		Diabetes		Kidney	
		Tests	Accuracy	Tests	Accuracy	Tests	Accuracy
4O-MINI	0.3	2.36 ± 0.68	0.83 ± 0.37	2.40 ± 0.73	0.58 ± 0.49	1.60 ± 0.75	1.00 ± 0.05
	0.5	1.87 ± 0.77	0.84 ± 0.37	2.06 ± 0.75	0.58 ± 0.49	1.48 ± 0.68	1.00 ± 0.04
	0.7	1.48 ± 0.73	0.85 ± 0.36	1.85 ± 0.73	0.58 ± 0.50	1.28 ± 0.52	1.00 ± 0.05
4O	0.3	2.67 ± 0.60	0.84 ± 0.37	2.25 ± 0.81	0.65 ± 0.48	1.35 ± 0.64	1.00 ± 0.04
	0.5	2.41 ± 0.67	0.83 ± 0.38	1.90 ± 0.88	0.67 ± 0.47	1.09 ± 0.35	1.00 ± 0.06
	0.7	2.27 ± 0.65	0.82 ± 0.39	1.67 ± 0.83	0.67 ± 0.47	1.04 ± 0.24	1.00 ± 0.06

Feature selection impact on performance

This section investigates the impact of feature selection frameworks on diagnostic performance by analysing the features most frequently selected and their empirical informativeness. Using a random baseline to mitigate confounding, we first identified empirically informative features based on their accuracy when selected among three features. For GPT-4o-mini, the most informative features were GGT, AST, and ALT for hepatitis; BloodPressure, Pregnancies, and Insulin for diabetes; and Random Blood Glucose, Potassium Levels, and Serum Creatinine for CKD. For GPT-4o, BIL, CHE, and GGT were most informative for hepatitis; BloodPressure, Skin Thickness, and Insulin for diabetes; and Blood Glucose, Red Blood Cell Count, and Potassium Levels for CKD. We also identified globally optimal features selected by both models as Glucose, BMI, and Insulin for diabetes, ALT, AST, and BIL for hepatitis, and BloodPressure, Serum Creatine, and Haemoglobin for CKD.

We then examined the feature selection patterns of ACTMED and the implicit baseline. The random baseline served as a control, showing no significant selection bias. For hepatitis, ACTMED with GPT-4o-mini preferentially selected ALT, AST, and GGT, while with GPT-4o it favoured AST, GGT, and BIL. The implicit method selected ALT, AST, and GGT for GPT-4o-mini and ALT, AST, and BIL for GPT-4o. On the diabetes dataset, GPT-4o-mini with ACTMED selected Glucose, Blood Pressure, and Skin Thickness, whereas GPT-4o preferred Glucose, Diabetes Pedigree Function, and BMI. The implicit method consistently selected BMI, Glucose, and Insulin across both models. For the CKD dataset, GPT-4o-mini with ACTMED selected Red Blood Cell Count, Blood Pressure, and Creatinine, while GPT-4o selected Blood Urea, Creatinine, and Blood Pressure. The implicit method selected Serum Creatinine, Blood Urea, and Blood Pressure for GPT-4o-mini, and Serum Creatinine, Blood Urea, and Haemoglobin for GPT-4o.

Collectively, these results indicate that neither ACTMED nor the implicit method consistently selects a fixed set of features that are empirically superior across all scenarios. The observed performance advantage of ACTMED likely stems not from identifying a universally optimal feature subset, but rather from its ability to perform more varied and potentially better-personalized test selections compared to the more constrained implicit or global methods, thereby optimizing the diagnostic process for individual cases.

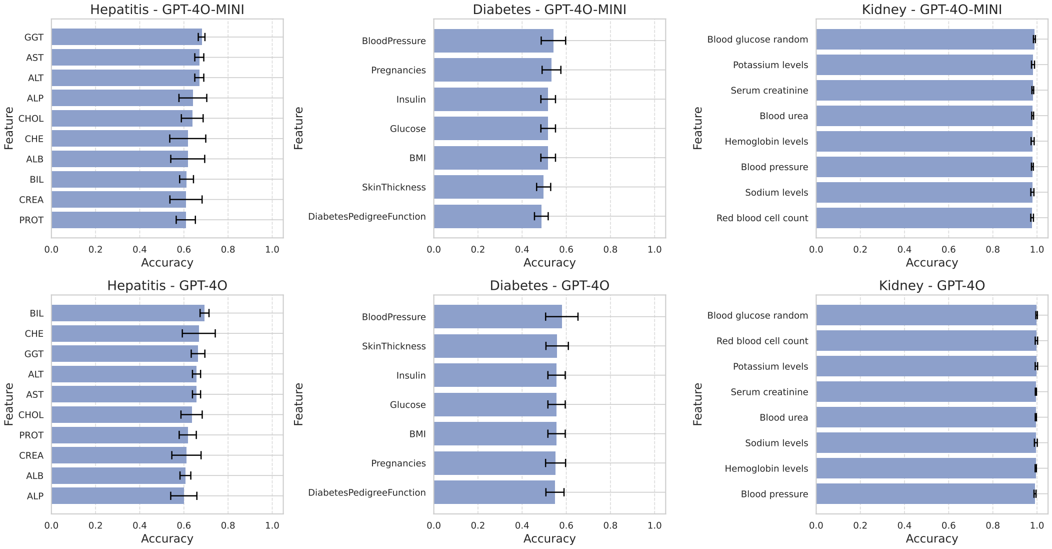

Figure 14: *Feature performance.* Mean accuracy per feature, measured across multiple seeds, for two models (gpt-4o-mini and gpt-4o) on three datasets. Bars represent mean accuracy (0–1), and black error bars denote ±1 standard deviation. Features are ranked in descending order of mean accuracy.

Example of ACTMED's Test Evaluation Process with Clinician-in-the-Loop

To illustrate ACTMED's information-theoretic test selection process, we present a simplified synthetic binary diagnostic task involving chronic kidney disease (CKD) with two tests. At each step, ACTMED supports clinician decision-making by maintaining transparency over test evaluations and allowing human review.

Step 1: Prior Belief. The system starts with a diagnostic prior based on available information:

$$P(y_{d_t} = 1 \mid K_t) = \mathbb{B}(p_{\text{prior}}), \quad \text{where} \quad p_{\text{prior}} = 0.20$$

Clinician role: The clinician can inspect the current risk estimate and adjust the prior based on additional context not currently captured in the structured data.

Step 2: Candidate Test Outcomes. ACTMED considers two candidate tests and simulates plausible results using a surrogate model:

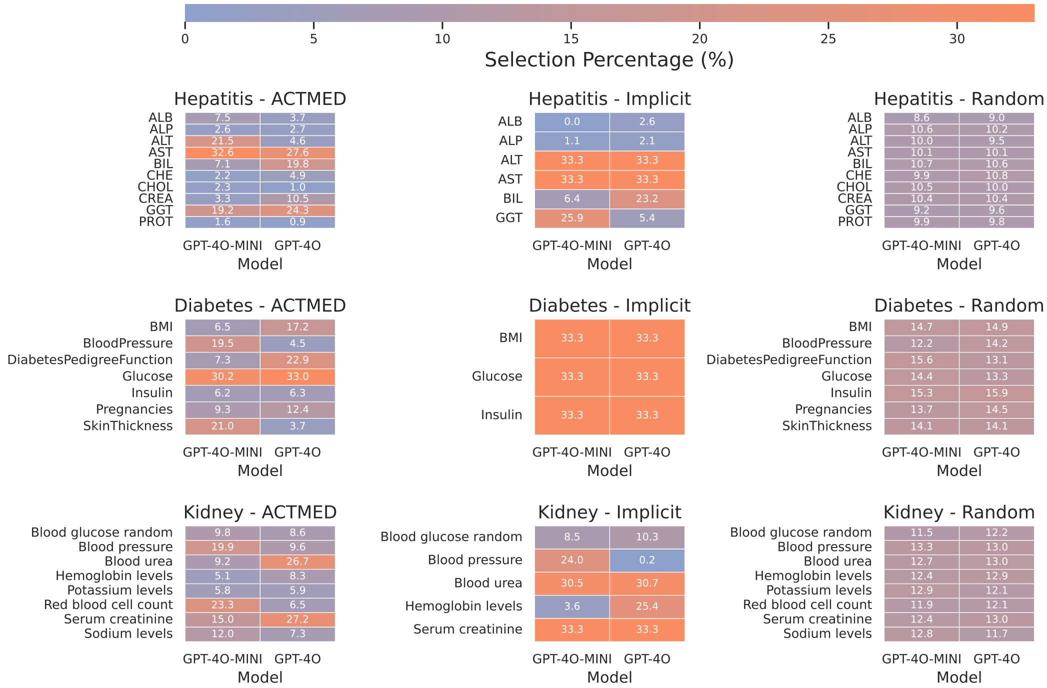

Figure 15: *Feature selection by method.* Heatmaps of feature-selection frequencies (0–100%) across three datasets (hepatitis, diabetes, and kidney) and three selection methods (ACTMED, Implicit, Random). Columns represent GPT-4O-Mini and GPT-4O. Each cell shows how often a feature was included across all seeds

- **Creatinine (high)** : 2.3, mg/dL
- **Creatinine (normal)** : 1.0, mg/dL
- **Sodium (normal)** : 140 mmol/L
- **Sodium (low)** : 130 mmol/L

Clinician role: The clinician may review the simulated outcomes for plausibility, reject irrelevant or infeasible tests, and flag preferred tests based on domain knowledge or patient-specific factors.

Step 3: Posterior Beliefs. ACTMED computes updated diagnostic beliefs for each hypothetical outcome:

- **Creatinine**
  - high: $P(y_{d_t} = 1 \mid K_t, \text{high}) = 0.65$
  - normal: $P(y_{d_t} = 1 \mid K_t, \text{normal}) = 0.22$
- **Sodium**
  - low: $P(y_{d_t} = 1 \mid K_t, \text{low}) = 0.45$
  - normal: $P(y_{d_t} = 1 \mid K_t, \text{normal}) = 0.18$

Clinician role: The clinician can examine the impact of each test result on the diagnostic belief, and assess whether these posterior shifts are clinically meaningful or likely to influence treatment decisions.

Step 4: Utility of Each Test. We compute the expected information gain from each test using the KL divergence between posterior and prior diagnostic beliefs. This utility is expressed as:

$$\mathcal{F}(u_t^{\text{crea}}) = \mathbb{E}[\text{KL}(P(y_{d_t} = 1 \mid K_t, u_t^{\text{crea}}) \parallel P(y_{d_t} = 1 \mid K_t))] = 0.134,$$

$$\mathcal{F}(u_t^{\text{sod}}) = \mathbb{E}[\text{KL}(P(y_{d_t} = 1 \mid K_t, u_t^{\text{sod}}) \parallel P(y_{d_t} = 1 \mid K_t))] = 0.056.$$

Since $\mathcal{F}(u_t^{\text{crea}}) > \mathcal{F}(u_t^{\text{sod}})$, ACTMED selects the serum creatinine test as it provides higher expected diagnostic value.

Clinician role: Before confirming the selected test, the clinician may override the choice if the utility estimate contradicts clinical judgment, safety concerns, or logistical constraints.

Once a test is acquired, the diagnostic process iteratively continues until a diagnosis is achieved. Figure 16 details how the model supports this by generating intermediate outputs that clinicians can review throughout the process.

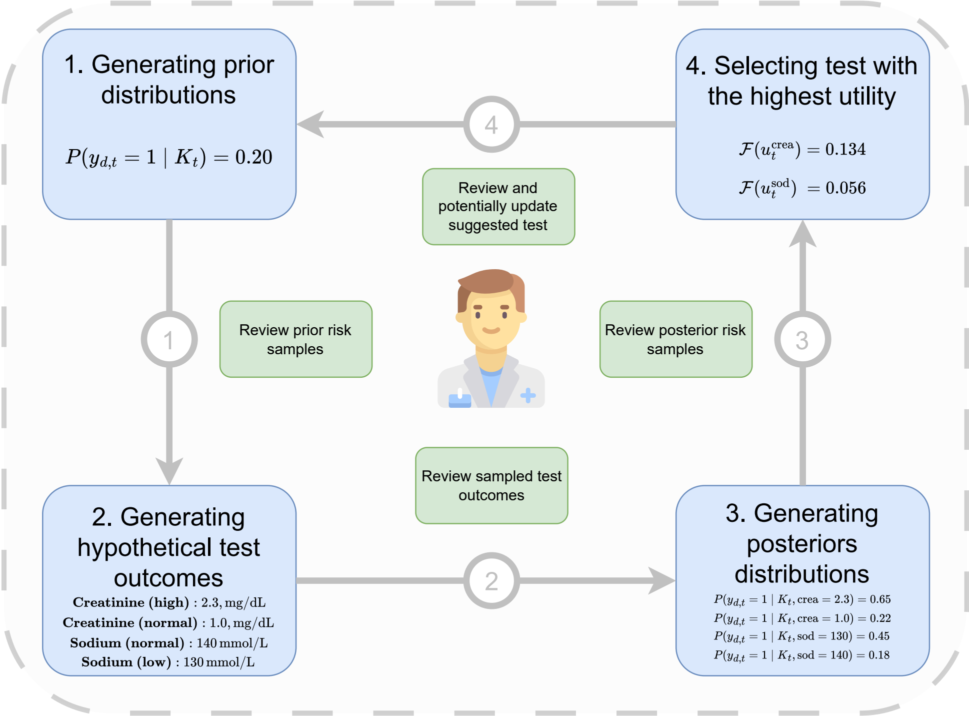
1. Generating prior distributions
$P(y_{d,t} = 1 \mid K_t) = 0.20$
4. Selecting test with the highest utility
$\mathcal{F}(u_t^{\text{crea}}) = 0.134$
$\mathcal{F}(u_t^{\text{sod}}) = 0.056$
4
Review and potentially update suggested test
1
Review prior risk samples
Review posterior risk samples
3
Review sampled test outcomes
2. Generating hypothetical test outcomes
Creatinine (high) : 2.3, mg/dL
Creatinine (normal) : 1.0, mg/dL
Sodium (normal) : 140 mmol/L
Sodium (low) : 130 mmol/L
2
3. Generating posteriors distributions
$P(y_{d,t} = 1 \mid K_t, \text{crea} = 2.3) = 0.65$
$P(y_{d,t} = 1 \mid K_t, \text{crea} = 1.0) = 0.22$
$P(y_{d,t} = 1 \mid K_t, \text{sod} = 130) = 0.45$
$P(y_{d,t} = 1 \mid K_t, \text{sod} = 140) = 0.18$

Figure 16: *Illustrative example of ACTMED's diagnostic reasoning.* The current belief (prior) about CKD is updated based on hypothetical outcomes for two candidate tests (serum creatinine and sodium). Posterior probabilities differ for each outcome, and the expected KL divergence determines which test offers the greatest diagnostic value. Serum creatinine yields higher expected information gain and is selected.

F Computational Cost Analysis

Our framework relies on repeatedly querying a LLM to simulate plausible estimates for candidate diagnostic test results and estimate disease posteriors. Consequently, the computational cost is dominated by the number of LLM queries required per patient episode.

At each decision step t, the agent evaluates a set of candidate diagnostic tests $U_t \subset \mathcal{U}_t$, where $|U_t|$ denotes the number of available tests. For each candidate test $u_t^{(i)} \in U_t$, the agent samples M possible outcomes and the resulting hypothetical posterior probability from the LLM to approximate the expected KL divergence. Thus, the computational cost per decision step is $\mathcal{O}(|U_t|M)$ LLM queries.

Let T denote the maximum number of decision steps per patient before either termination or diagnosis. Then, the total computational cost per patient is:

$$\mathcal{C} = \mathcal{O}\left(\sum_{t=1}^{T} |U_t| M\right). \tag{9}$$

In the worst case, where no early termination occurs and all tests are considered at every step, the complexity simplifies to:

$$\mathcal{C} = \mathcal{O}(TNM), \tag{10}$$

where N is the total number of possible diagnostic tests. In practice, N may already be quite small as there will only be a subset of all available tests that can be ordered for diagnosing conditions due to existing guidelines and the model only needs to determine which of those offers the highest utility. Table 9 summarizes the asymptotic computational complexity of our method compared to baseline approaches. While our method incurs higher per-patient computational cost compared to static classifiers, this is justified in domains like clinical medicine where information acquisition is expensive and decision quality is paramount. Inference time depends primarily on LLM latency and hardware availability. In our setup, estimating the expected value of a single diagnostic test takes approximately 20 seconds using either GPT-4o or GPT-4o-mini. Running the full suite of experiments across five random seed, parallelized under a shared API key for each model, takes roughly 60 hours. In deployment, inference time could be significantly reduced through parallelization of the independent API requests. This cost and time delay is negligible relative to the clinical and financial burden of unnecessary or delayed testing. Furthermore, computational demands can be significantly reduced by training lightweight surrogate models specialized for predicting test outcomes, as demonstrated in prior work [82].

Table 9: Comparison of per-patient computational complexity across methods.

Method	**Complexity**	**Description**
Static classifier	$\mathcal{O}(1)$	Single forward pass using all features
Stochastic feature acquisition	$\mathcal{O}(1)$	Random subset of all features selected
Greedy feature acquisition	$\mathcal{O}(T)$	Selects top-k features in static order
BED with KL divergence (ours)	$\mathcal{O}(TNM)$	Actively selects using KL divergence