BAYESIAN AGENTIC MEDICAL DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient sequential clinical diagnosis typically requires maintaining calibrated beliefs over diseases and selecting tests that maximally resolve uncertainty. Current LLM-based diagnostic agents handle both tasks inside the model itself, through accumulated context, verbalized confidence, or prompt sampling, inheriting the LLM's miscalibration and discarding the principled structure of Bayesian decision theory. We revisit this architectural miscalibration and introduce Bayesian Agentic Medical Diagnosis, abbreviated as "BAMD", which confines the LLM to a predictive role while delegating belief maintenance and test selection to an external Bayesian engine. Beliefs are updated via Bayes' rule after each observation, and test selection is governed by information-theoretic operators on the evolving belief state. On four clinical benchmarks, BAMD integrated on an 8B Bayesianized Llama model running on a single local GPU matches or exceeds prompt sampling with a locally deployed 120B LLM across all four datasets, while producing the best-calibrated probability estimates of any LLM-based approach we compare against. These results highlight that the limiting factor in LLM-based sequential diagnosis is architectural, not scale-related, and that a small model embedded in the proper framework can outperform much larger proprietary systems accessed through commercial APIs.

1 INTRODUCTION

Clinical diagnosis is inherently sequential (Wald, 1992). A physician begins with a limited patient profile, forms initial hypotheses, and iteratively orders tests whose results sharpen or redirect those hypotheses until a confident diagnosis emerges (Sox et al., 2024). Every test carries a cost: financial, temporal, and patient discomfort. The physician's core challenge is therefore not just prediction but also *control*: deciding which test to order next, and when the accumulated evidence is sufficient to stop. Formalizing this control problem requires maintaining an explicit belief state over the disease hypothesis, updating it as new evidence arrives, and comparing the expected value of acquiring additional information against its cost. Together, these components constitute the classical machinery of Bayesian decision theory (Berger, 2013; DeGroot, 2005), studied extensively under sequential analysis (Wald, 1992) and Bayesian experimental design (Lindley, 1956; Chaloner & Verdinelli, 1995); a body of work that long predates the arrival of large language models (LLMs).

Recently, LLMs have demonstrated strong reasoning performance across medical settings, with profound benefits on licensing examinations (Nori et al., 2023; Singhal et al., 2023), generating clinically plausible differential diagnoses (McDuff et al., 2025), and conducting multi-turn diagnostic dialogues that rival primary care physicians (Tu et al., 2024). This naturally motivates the deployment of LLMs in the sequential diagnostic workflow, where the central challenge shifts from processing a fixed input to actively deciding which information to acquire next. Recently, this direction has been pursued along two broad lines in literature: (i) Surrogate-based Bayesian experimental design uses the LLM as a conditional sampler to simulate plausible test outcomes and selects the next test through the expected information gain (Estévez et al., 2025); and (ii) Reinforcement learning-based approaches train LLM-based agents end-to-end to emit verbalized confidence scores and select tests via prompted deliberation (Bani-Harouni et al., 2025; Sun et al., 2024; Denton, 2018). Additionally, prompted multi-agent orchestration coordinates off-the-shelf LLMs through structured deliberation (Nori et al., 2025). Clinical deployment also places practical constraints on where patient data can be processed. Large clinical resources such as MIMIC-IV are distributed under controlled-access procedures (Johnson et al., 2023), making effective local inference particularly attractive for data-sensitive diagnostic systems.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Underlying these paradigms, there is a shared architectural assumption whose implications for sequential diagnostic performance are scrutinized in this paper. In these paradigms, the belief state is placed inside the LLM itself, either as the accumulating content of a context window, a verbalized integer confidence score, or a distribution induced by prompt sampling. In addition, they delegate test selection and stopping criteria to the LLM's internal outputs, rather than to an explicit decision-theoretic mechanism. The resulting LLM-based agents inherit two well-documented and scale-invariant failure modes of LLMs: (i) *miscalibrated* probability estimates for disease prediction that remain unreliable even in state-of-the-art models and are difficult to refine through prompting alone (Savage et al., 2024; Xiong et al., 2023; Kapoor et al., 2024), and (ii) *hallucinated* clinical reasoning that yields confidently incorrect outputs without a verifiable external record of how evidence was accumulated (Hager et al., 2024; Pal et al., 2023). Figure 1 illustrates the empirical impact of these failure modes on diagnostic performance across four clinical datasets. In sequential diagnosis, where each decision depends critically on the probability estimate from the previous step and informs subsequent test selection based on accumulated context, these failure modes propagate rather than remain isolated: miscalibration accumulates, inducing premature termination and systematic misselection of tests, while hallucinated intermediate reasoning progressively distorts the belief state away from the ground-truth signal that would be anchored by actual test results. The core design principles of sequential diagnosis, namely explicit belief tracking, information-theoretic test selection, and principled stopping, remain unchanged in the LLM setting; what differs is the architectural machinery required to realize them.

Driven by these insights, we propose an architectural alternative for LLM-based agents. We introduce **BAMD** (Bayesian Agentic Medical Diagnosis), a framework that reinstates classical sequential decision-making machinery while restricting the LLM to a predictive role within it. Specifically, a Bayesianized LLM, obtained via variational fine-tuning of low-rank adapters (Wang et al., 2024), provides posterior predictive distributions over disease risk and test-outcome likelihoods. An explicit belief state is maintained *outside* the LLM, initialized from the LLM's posterior predictive distribution introduced above and updated via Bayes' rule as test results arrive. Grounded in information-theoretic first principles, a control layer selects the next test by maximizing expected information gain per unit cost and determines when to stop using a value-of-information criterion. The LLM neither maintains beliefs, selects tests, nor determines when to stop. It produces calibrated predictive distributions, with all downstream decisions delegated to an *external* Bayesian controller.

By placing belief and control outside the LLM, we fundamentally expand what is structurally possible: (i) The belief state becomes transparent and auditable across decision steps instead of being absorbed into context; (ii) Posterior averaging over weight samples yields probability estimates whose magnitudes are semantically meaningful, rather than artifacts of token generation; (iii) Value-of-information stopping becomes a principled operation on a continuous belief, rather than a heuristic over discretized outputs; and (iv) the entire system operates on a single consumer-grade GPU, ensuring that patient data remain local in environments where privacy, regulatory, or infrastructure constraints preclude reliance on frontier-model APIs. Figure 3 in the supplementary file contrasts the representative architectural paradigms for LLM-based sequential diagnosis and highlights where belief maintenance and decision control reside in each design.

Contributions. The contributions of this paper can be contextualized in the following aspects: **First**, we argue and formalize that LLM-based sequential diagnostic agents should not maintain internal beliefs. Rather, belief must be externalized as a Bayesian agent and updated through an explicit Bayes-rule operator using calibrated likelihoods supplied by the LLM. **Second**, we argue and formalize that such an agent should not control its own decision process. Test selection and stopping must be delegated to information-theoretic operators over the external belief, rather than relying on the LLM's outputs.**Third**, we demonstrate empirically that a Bayesianized 8B Llama model, embedded within this external Bayesian framework, *matches or exceeds prompt-based sampling with a locally deployed 120B LLM* (Estévez et al., 2025) on every clinical dataset we evaluate. This includes a direct comparison under the same evaluation protocol, where the 8B local model outperforms the 120B local model in AUC, accuracy, and F1 across all four datasets. Across four clinical benchmarks, BAMD yields the best-calibrated probability estimates among all LLM-based counterparts we evaluate, while operating entirely on a single consumer-grade GPU with patient data remaining on local hardware.

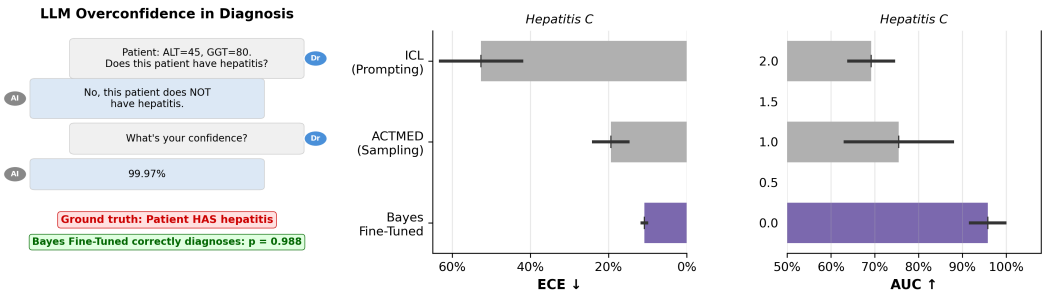


Figure 1: Empirical illustration of two LLM failure modes (miscalibration and hallucination) in sequential clinical diagnosis, averaged across four clinical datasets (CKD, hepatitis, heart disease, diabetes). All three agents share the same Llama8B backbone and differ only in their paradigm-specific inference procedures: ICL (prompting) and ACTMED (outcome-level sampling), described in Section B, and the proposed BAMD (Bayesian fine-tuning) presented in Section 3. The ECE subplot (↓ better) reports expected calibration error; the AUC subplot (↑ better) reports diagnostic discrimination. The two prompt-based paradigms exhibit substantially higher ECE and lower AUC than BAMD, empirically reflecting the impact of the miscalibration and hallucination failure modes discussed above when belief tracking and decision-making are confined inside the LLM.

2 Problem Formulation

Consider that a patient is characterized by a demographic feature vector $\mathbf{x}$ and is subject to a shared panel of M candidate clinical tests $\mathcal{T} = \{T_1, \ldots, T_M\}$ indexed by $j \in \{1, \ldots, M\}$, each with a known acquisition cost $c_j > 0$. Let $D \in \mathcal{D}$ denote the *unknown* disease status of the patient, and let $Z_j \in \mathcal{Z}$ denote the outcome of test T_j, with lowercase d and z_j denoting their realized values. For brevity, we write $p(d \mid \cdots) \equiv p(D = d \mid \cdots)$ and $p(z \mid d, \cdots) \equiv p(Z_j = z \mid D = d, \cdots)$ throughout. In the binary diagnosis setting considered throughout this paper, $\mathcal{D} = \{0, 1\}$ encodes (healthy, diseased) and $\mathcal{Z} = \{0, 1\}$ encodes (normal, abnormal).

A diagnostic episode for the patient proceeds over discrete turns $t = 1, 2, \ldots$. At the start of turn t, the agent has recorded the ordered-test index set up to turn $t-1$, $\mathcal{S}_{t-1} = \{j_1, j_2, \ldots, j_{t-1}\} \subseteq \{1, \ldots, M\}$, where $j_\tau \in \{1, \ldots, M\}$ denotes the test index selected at turn τ for the patient, together with the corresponding observation set $\mathcal{O}_{t-1} = \{(j, z_j) : j \in \mathcal{S}_{t-1}\}$, in which $z_j \in \mathcal{Z}$ is the realized outcome of test T_j. The agent maintains a disease belief

$$r_t(d) \triangleq p\big(D = d \,\big|\, \mathbf{x}, \mathcal{O}_{t-1}\big), \quad d \in \mathcal{D}. \tag{1}$$

At each turn, the agent either selects a new test index $j_t \in \{1, \ldots, M\} \setminus \mathcal{S}_{t-1}$, applies test T_{j_t}, and observes outcome $z_{j_t} \in \mathcal{Z}$, or terminates the episode at stopping time τ and reports the maximum-a-posteriori diagnosis

$$\hat{d} = \arg\max_{d \in \mathcal{D}} r_\tau(d). \tag{2}$$

In addition, we consider an upper-bound $B \geq 1$ to the number of ordered tests per patient; that is $|\mathcal{S}_\tau| \leq B$. The agent's objective is to maximize diagnostic accuracy while minimizing the total acquisition cost $\sum_{j \in \mathcal{S}_\tau} c_j$.

Existing LLM-based approaches for sequential diagnosis. Two LLM-based paradigms have been explored in the literature for this problem. *In-context learning (ICL)* (Nori et al., 2025) prompts the LLM at each turn to name the next test and read off the final diagnosis, with the belief state implicit in the accumulating context. *ACTMED* (Estévez et al., 2025) retains the classical Bayesian experimental design objective but uses the LLM as a zero-shot conditional outcome sampler, scoring tests by the empirical expected KL divergence between posterior and prior risk. Please refer to Appendix B for the details of these two approaches. Figure 1 summarizes the empirical behavior of these two paradigms against a Bayesian fine-tuning the alternative that we develop in Section 3, averaged across the four clinical datasets introduced in Section 4.1. Both ICL and ACTMED produces diagnostically informative decisions on average, but neither closes the calibration gap: ICL's context-based belief drifts as the vignette accumulates, and ACTMED's outcome-level sampling distribution collapses when the underlying fine-tuned surrogate is sharply peaked. These failure modes are structural rather than incidental, motivating the architectural shift we formalize in the remainder of this paper. Neither paradigm maintains an external belief state r_t as defined in Eq. equation 1 that can be updated via

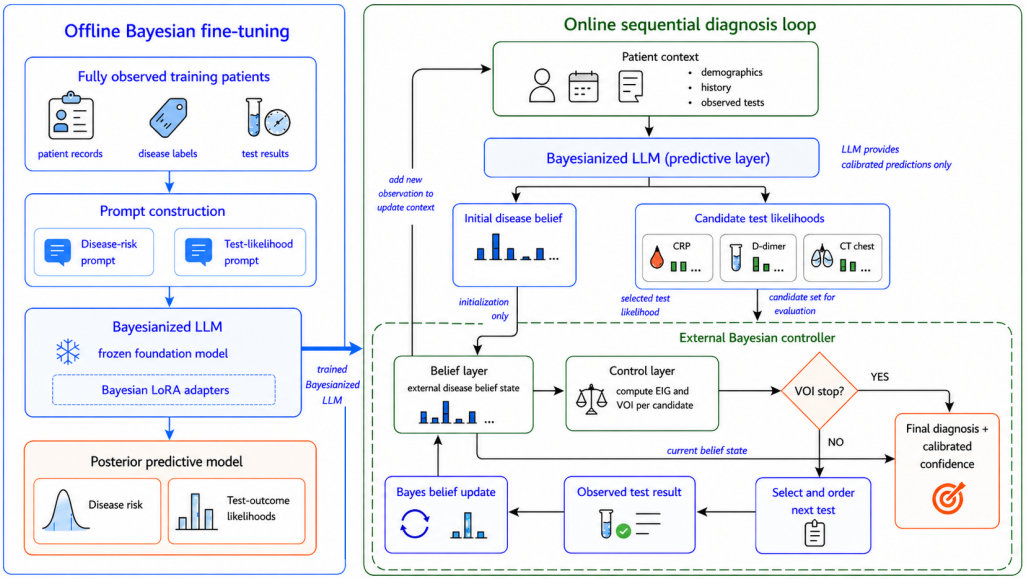


Figure 2: Overview of the proposed Bayesian Agentic Medical Diagnosis (BAMD) framework. Left: offline Bayesian fine-tuning on fully observed training patients, where disease-risk and test-likelihood prompts are used to train a Bayesianized LLM with Bayesian LoRA adapters, yielding a posterior predictive model over disease risk and test-outcome likelihoods. Right: online sequential diagnosis loop, where the Bayesianized LLM provides calibrated predictive signals only, while an external Bayesian controller maintains the disease belief state, computes expected information gain and value of information, selects the next test, updates the belief via Bayes' rule after observing test outcomes, and outputs the final diagnosis with calibrated confidence.

Bayes rule, carried across turns, and coupled with a value-of-information stopping criterion. Building this external state requires the LLM to emit calibrated likelihood signals so that the downstream Bayesian updates are reliable, which we address through the Bayesian fine-tuning process described in Section 3.1.

3 METHOD

In this section, we present the Bayesian Agentic Medical Diagnosis (BAMD) framework. BAMD structures sequential diagnosis into three layers that operate in a closed loop at each turn. The *predictive layer* is the Bayesian fine-tuned LLM, which supplies posterior predictive distributions over the disease status and over the outcome of each candidate test conditioned on a hypothesized disease, with parameter uncertainty propagated through the variational posterior q_ϕ. The *belief layer* maintains the disease posterior $r_t(d) = p(D = d \mid \mathbf{x}, \mathcal{O}_{t-1})$ of Eq. equation 1 externally to the LLM, initializing it from the predictive layer and updating it recursively via Bayes' rule as test outcomes arrive on-the-fly. The *control layer* selects the next test by maximizing expected information gain per unit cost, and terminates testing by comparing the value of further information against its cost.

We posit that the core quantities underpinning principled sequential diagnosis, namely disease posterior, mutual information between prospective test outcomes and disease, expected utility under the current belief, and the value of additional information, are all exactly computable from the posterior predictive distributions of the Bayesianized LLM. This design removes sequential-decision responsibilities from the LLM and delegates them to external modules that operate deterministically on its calibrated predictive outputs. This separation distinguishes BAMD from the in-context and prompt-sampling paradigms reviewed in App. B, which conflate the predictive model with the decision-maker.

3.1 BAYESIAN FINE-TUNING OF LLMS

Let $\{(\mathbf{x}_n, D_n, \{z_{n,j}\}_{j=1}^{M})\}_{n=1}^{N_{\text{tr}}}$ denote a set of N_{tr} fully-observed training patients with known disease status D_n and known test outcomes $z_{n,j}$ for every test $T_j \in \mathcal{T}$. From these patients we

construct a supervised prompt-answer dataset $\mathcal{D}_{\text{tr}} := \{(\boldsymbol{\varphi}_i,\, y_i)\}_{i=1}^{N_q}$ (please refer to App. C in the supplementary file for its construction), based on which we will fine-tune the LLM. For parameter-efficient fine-tuning, we will adopt the low-rank adaptation (LoRA) (Hu et al., 2022), which, given pre-trained LLM weights $\mathbf{W}_0 \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$, yields fine-tuned weights given by $\mathbf{W} := \mathbf{W}_0 + \mathbf{BA}$, where $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{out}}}$ and $\mathbf{B} \in \mathbb{R}^{d_{\text{in}} \times r}$ are two low-rank matrices with rank $r \ll \{d_{\text{in}}, d_{\text{out}}\}$. To further endow fine-tuned LLM with uncertainty quantification, we consider *random* $\mathbf{A}$ with prior given by an isotropic Gaussian $p(\mathbf{A}) = \mathcal{N}(0,\, \beta^2\mathbf{I})$ (the scale $\beta > 0$ is pre-selected), while keeping $\mathbf{B}$ *deterministic*. This asymmetric design, inherited from (Wang et al., 2024) for scalability, exploits the observation that placing a Gaussian on one factor of $\mathbf{A}$ already captures the dominant directions of weight uncertainty (Wang et al., 2024).

Our goal is to jointly find the posterior of $\mathbf{A}$ along with the point estimate of $\mathbf{B}$. For scalability, we will leverage the variational inference, where we place a factorized Gaussian variational posterior on each LoRA-A matrix, $q(\mathbf{A}) = \mathcal{N}(\boldsymbol{\mu},\, \text{diag}(\boldsymbol{\sigma}^2))$ with *learnable* mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma} = (\boldsymbol{\rho})^2$. Collecting all the learnable parameters $\boldsymbol{\phi} := \{\boldsymbol{\mu},\, \boldsymbol{\rho},\, \mathbf{B}\}$, fine-tuning is carried out by minimizing the negative evidence lower bound (ELBO)

$$\mathcal{L}_{\text{ELBO}}(\boldsymbol{\phi}) \;=\; -\,\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{A})}\Big[\log p(\mathcal{D}_{\text{tr}} \mid \mathbf{A}, \hat{\mathbf{B}})\Big] \;+\; \pi_t \cdot \text{KL}\big(q_{\boldsymbol{\phi}}(\mathbf{A}) \,\big\|\, p(\mathbf{A})\big)\,, \tag{3}$$

where $\pi_t \in (0, 1]$ is the minibatch-dependent KL reweighting coefficient of (Blundell et al., 2015), and we estimate the expectation via local reparameterization with flipout (Kingma et al., 2015; Wen et al., 2018).

Having obtained the estimated $\hat{\mathbf{B}}$ and $q(\mathbf{A})$, we will rely on Monte-Carlo sampling to obtain the following posterior predictive distribution for any new query $\mathbf{q}$, by drawing K posterior samples $\tilde{\mathbf{A}}^{(k)} \sim \mathcal{N}(\hat{\boldsymbol{\mu}},\, \text{diag}(\hat{\boldsymbol{\sigma}}^2))$, namely

$$p(y \mid \mathbf{q},\, \mathcal{D}_{\text{tr}}) \approx \int p^{\text{LLM}}(y \mid \mathbf{q}, \mathbf{A}, \hat{\mathbf{B}})\, q_{\boldsymbol{\phi}}(\mathbf{A})\, d\mathbf{A} \approx \; \frac{1}{K}\sum_{k=1}^{K} p^{\text{LLM}}\Big(y \,\Big|\, \mathbf{q}, \tilde{\mathbf{A}}^{(k)}, \hat{\mathbf{B}}\Big) \tag{4}$$

which forms the core prediction principle in the proposed BAMD framework, as will be shown subsequently.

3.2 Belief propagation

At the start of turn $t = 1$ no test has yet been ordered and $\mathcal{O}_0 = \emptyset$, so the initial belief r_1 depends only on the demographic features $\mathbf{x}$. We define r_1 as the posterior predictive disease probability obtained by marginalizing the LLM's output over the variational posterior $q_{\boldsymbol{\phi}}$,

$$r_1(d) \;=\; p(d \mid \mathbf{x}) \;=\; \int p^{\text{LLM}}(d \mid \mathbf{x}, \mathbf{A}, \hat{\mathbf{B}})\, q_{\boldsymbol{\phi}}(\mathbf{A})\, d\mathbf{A} \approx \frac{1}{K}\sum_{k=1}^{K} p^{\text{LLM}}\Big(d \,\Big|\, \mathbf{x}, \tilde{\mathbf{A}}^{(k)}, \hat{\mathbf{B}}\Big) \tag{5}$$

where $p^{\text{LLM}}(d \mid \mathbf{x}, \mathbf{A}, \hat{\mathbf{B}})$ is the Yes-token probability produced by the LLM under parameter setting $(\mathbf{A}, \hat{\mathbf{B}})$ on the Type B prompt of Section C.

3.3 Cost-aware test selection

At turn $t \geq 1$ with current belief r_t, we select the next test by maximizing the expected information gain (EIG) per unit acquisition cost. Computing the EIG requires two posterior predictive quantities defined against the variational posterior $q_{\boldsymbol{\phi}}$. The first is the class-conditional outcome likelihood

$$p_{j,t}(z \mid d) = \int p^{\text{LLM}}(z \,|\, d,\, \mathbf{x},\, \mathcal{O}_{t-1}, \mathbf{A}, \hat{\mathbf{B}})\, q_{\boldsymbol{\phi}}(\mathbf{A})\, d\mathbf{A} = \frac{1}{K}\sum_{k=1}^{K} p^{\text{LLM}}\Big(z \,\Big|d,\, \mathbf{x},\, \mathcal{O}_{t-1}, \tilde{\mathbf{A}}^{(k)}, \hat{\mathbf{B}}\Big) \tag{6}$$

for $j \in \mathcal{J}_{t-1}$, $d \in \mathcal{D}$, and $z \in \mathcal{Z}$, where the integrand is obtained by feeding the Type A prompt of Section C conditioned on hypothesis d to the LLM and reading the Yes-token probability. The integral in Eq. equation 6 is approximated by a Monte Carlo estimator using the same K posterior samples drawn for Eq. equation 5,

Given equation 6, the marginal predictive outcome probability under the current disease belief is given by

$$p_{j,t}^{\text{mar}}(z) \;=\; \sum_{d\in\mathcal{D}} r_t(d)\, p_{j,t}(z \mid d), \qquad z \in \mathcal{Z}. \tag{7}$$

We define the EIG of candidate test T_j at turn t as the mutual information between its prospective outcome Z_j and the disease status D under the current vignette,

$$\text{EIG}_t(j) \;:=\; I\big(Z_j;\, D \,\big|\, \mathbf{x},\, \mathcal{O}_{t-1}\big) \;=\; H\big[p_{j,t}^{\text{mar}}\big] \;-\; \sum_{d\in\mathcal{D}} r_t(d)\, H\big[p_{j,t}(z \mid d)\big]\,, \tag{8}$$

where $H[\cdot]$ is the Shannon entropy of a discrete distribution over $\mathcal{Z}$. The selected test index is

$$j_t \;=\; \arg\max_{j\in\mathcal{J}_{t-1}} \frac{\text{EIG}_t(j)}{c_j}. \tag{9}$$

3.4 Belief update

Once the agent applies test T_{j_t} and observes the realized outcome $z_t \in \mathcal{Z}$, we update the belief via Bayes' rule using the posterior predictive class-conditional likelihood of Eq. equation 6,

$$r_{t+1}(d) \;=\; \frac{r_t(d)\, p_{j_t,\,t}(z_t \mid d)}{\sum_{d'\in\mathcal{D}} r_t(d')\, p_{j_t,\,t}(z_t \mid d')} \qquad d \in \mathcal{D}, \tag{10}$$

and extend the ordered-test index set and the observation set as $\mathcal{S}_t \;=\; \mathcal{S}_{t-1} \cup \{j_t\}$ and $\mathcal{O}_t \;=\; \mathcal{O}_{t-1} \cup \{(j_t,\, z_t)\}$.

3.5 Value-of-information (VoI) stopping

Our BAMD framework features a novel VoI stopping mechanism inspired by (Papamarkou et al., 2024). Specifically, let $\mathcal{A}$ denote a finite set of downstream clinical actions, for example $\mathcal{A} = \{\text{treat},\, \text{no-treat}\}$ in the binary case, and let $u : \mathcal{A}\times\mathcal{D} \to \mathbb{R}$ be a utility function with $u(a, d)$ denoting the utility of taking action a when the true disease status is $D = d$. The expected utility of action a under a belief r is

$$\text{EU}(a \mid r) \;=\; \sum_{d\in\mathcal{D}} r(d)\, u(a, d), \qquad a \in \mathcal{A}, \tag{11}$$

and we write $\text{EU}^*(r) := \max_{a\in\mathcal{A}} \text{EU}(a \mid r)$ for the optimal expected utility under r. We define the value of information (VOI) of a candidate test T_j at turn t as the expected improvement in optimal utility after observing its hypothetical outcome,

$$\text{VOI}_t(j) \;:=\; \sum_{z\in\mathcal{Z}} p_{j,t}^{\text{mar}}(z)\; \text{EU}^*\Big(r_{t+1}^{[z]}\Big) \;-\; \text{EU}^*(r_t), \qquad j \in \mathcal{J}_{t-1}, \tag{12}$$

where $r_{t+1}^{[z]}$ is the hypothetical belief that would result from applying Eq. equation 10 after observing outcome z on test T_j, and the outcomes are weighted by the marginal predictive in Eq. equation 7. Testing stops at turn t whenever no remaining test offers net positive expected value,

$$\max_{j\in\mathcal{J}_{t-1}} \big[\text{VOI}_t(j) - c_j\big] \;<\; 0, \tag{13}$$

in which case we set the stopping time $\tau = t$ and report the final diagnosis $\hat{d} = \arg\max_{d\in\mathcal{D}} r_\tau(d)$ as in Eq. equation 2.

Please refer to Fig. 2 for the flowchart of the proposed BAMD framework and Algs.1-2 in the supplementary files for the detailed implementations.

4 Experimental Results

4.1 Experimental setup

Datasets. We evaluate BAMD on four binary diagnosis datasets drawn from the UCI Machine Learning Repository (Asuncion et al., 2007): Chronic Kidney Disease (CKD) (Kalantar-Zadeh et al.,

2021), Hepatitis C (Bruggmann et al., 2014), Heart Disease (Cleveland) (Wingard et al., 1995), and Diabetes (Pima Indians) (Mousa et al., 2023). These cohorts span panel sizes from six to ten tests and diagnostic complexities ranging from a near-deterministic biomarker (CKD) to a class-imbalanced cohort with no single dominant biomarker (Diabetes). The Hepatitis C cohort uses the same 56-positive/56-healthy class balance as (Estévez et al., 2025), enabling identical-cohort comparison. All datasets use an 80/20 train/test split with a fixed seed. Per-dataset characteristics are provided in Appendix G.1.

Methods and baselines. Unless otherwise stated, the LLM-based methods use Llama-3.1-8B-Instruct (Grattafiori et al., 2024) with 8-bit quantization as the base model. BAMD, MLE, and MC-Dropout share the same LoRA configuration ($r = 8$) and training data, and differ only in the treatment of parameter uncertainty: a variational posterior on LoRA-A for BAMD, a point estimate for MLE, and inference-time dropout over the MLE point estimate for MC-Dropout. We further compare against the 8B reimplementation of ACTMED (Estévez et al., 2025) as the surrogate-based BED baseline, LA-CDM (Bani-Harouni et al., 2025) as the RL-based baseline, and an 8B adaptation of MAI-DxO (Nori et al., 2025) as the ICL-based baseline. We additionally instantiate ACTMED with NVIDIA Nemotron 3 Super 120B-A12B (NVIDIA, 2026). For broader context, we also report the ACTMED + GPT-4o and ACTMED + GPT-4o-mini results from (Estévez et al., 2025) on the three shared datasets (CKD, Hepatitis, and Diabetes). Full method configurations and hyperparameters are deferred to Table 3.

Evaluation protocol. At inference, BAMD draws $K = 10$ posterior weight samples per LLM query. All methods operate under a fixed test budget of $B = 3$ unless otherwise stated. We report diagnostic performance via AUC, accuracy, and F1, and calibration via ECE, negative log-likelihood, and Brier score. All results are averaged across 5 random seeds.

4.2 Diagnostic accuracy

Table 1 reports diagnostic performance over six LLM-based methods, a locally deployed 120B ACTMED instantiation, and two ACTMED instantiations with frontier APIs. BAMD's advantage over MLE, which constitutes the closest controlled comparator, sharing the same external controller and the same training data, increases monotonically with dataset difficulty: negligible on CKD, slight on hepatitis, more pronounced on heart disease, and greatest on diabetes. This pattern aligns with the theoretical role of posterior averaging, whose benefits are most pronounced in regimes where the likelihood remains genuinely uncertain rather than being dominated by a single, saturating biomarker.

Across the four datasets we evaluate, BAMD matches or outperforms ACTMED (120B), despite running an 8B model on a single local GPU rather than a substantially larger local model. On heart disease, where scaling ACTMED from 8B to 120B yields the largest AUC improvement, BAMD still outperforms ACTMED (120B) in AUC, accuracy, and F1. For additional context, BAMD also matches or outperforms the reported ACTMED + GPT-4o results on the three datasets shared with (Estévez et al., 2025). This indicates that the limiting factor in LLM-based sequential diagnosis is architectural rather than scale-related.

A backbone-agnostic check with Qwen-2.5-7B-Instruct (Table 7) preserves the same ranking, indicating that the gains stem from the architecture rather than from a specific backbone. Per-patient call complexity, wall-clock runtime on a single consumer GPU, and an API-cost comparison against baselines + GPT-4o are reported in Appendix G.4 Tables 8 and 9, showing that BAMD attains the above accuracy with competitive runtime on local hardware and no recurring per-query cost.

4.3 Calibration quality

Calibration metrics presented in Table 2 substantiate the empirical predictions of our proposed architectural design. MLE shares BAMD's external controller and differs only in whether the LoRA-A matrices carry a variational posterior. It produces NLL two to seven times worse and ECE up to four times worse than BAMD across the four datasets. The calibration gap exceeds the accuracy gap in relative terms, consistent with the theoretical role of posterior averaging: a point estimate may rank patients competitively, but it cannot produce probability values whose magnitudes are informative. In sequential diagnosis, where each step feeds on the previous step's probability through a Bayes-rule update and a value-of-information comparison, this calibration deficit accumulates.

Table 1: Diagnostic performance comparison (budget B=3, over 5 seeds). Best in **bold**.

		CKD	Hepatitis	Heart	Diabetes
AUC ↑	BAMD (ours, 8B local)	**1.000**$\pm$**.000**	**.975**$\pm$**.025**	**.907**$\pm$**.011**	.800$\pm$.025
	MC-Dropout (8B local)	1.000$\pm$.000	.923$\pm$.044	.833$\pm$.025	.721$\pm$.031
	MLE (8B local)	1.000$\pm$.000	.896$\pm$.004	.822$\pm$.024	.693$\pm$.025
	MAI-DxO (8B local)	1.000$\pm$.000	.692$\pm$.055	.716$\pm$.025	.730$\pm$.041
	ACTMED (8B local)	.983$\pm$.021	.755$\pm$.126	.544$\pm$.048	.635$\pm$.049
	LA-CDM (8B local)	.963$\pm$.036	.585$\pm$.151	.559$\pm$.011	.676$\pm$.002
	ACTMED + Nemotron-120B (local)	.997$\pm$.003	.780$\pm$.061	.791$\pm$.001	.677$\pm$.032
	ACTMED + GPT-4o (frontier API)	1.000$\pm$.000	.917$\pm$.013	—	**.803**$\pm$**.009**
	ACTMED + GPT-4o-mini (frontier API)	1.000$\pm$.000	.891$\pm$.006	—	.743$\pm$.009
Accuracy ↑	BAMD (ours, 8B local)	**1.000**$\pm$**.000**	**.957**$\pm$**.043**	**.860**$\pm$**.020**	**.740**$\pm$**.032**
	MC-Dropout (8B local)	1.000$\pm$.000	.899$\pm$.054	.796$\pm$.021	.697$\pm$.024
	MLE (8B local)	1.000$\pm$.000	.870$\pm$.022	.793$\pm$.007	.692$\pm$.011
	MAI-DxO (8B local)	1.000$\pm$.000	.667$\pm$.070	.700$\pm$.020	.710$\pm$.035
	ACTMED (8B local)	.979$\pm$.015	.797$\pm$.108	.560$\pm$.024	.626$\pm$.049
	LA-CDM (8B local)	.958$\pm$.015	.536$\pm$.160	.553$\pm$.016	.675$\pm$.033
	ACTMED + Nemotron-120B (local)	.958$\pm$.015	.739$\pm$.094	.717$\pm$.029	.569$\pm$.053
	ACTMED + GPT-4o (frontier API)	.999$\pm$.003	.839$\pm$.013	—	.668$\pm$.009
	ACTMED + GPT-4o-mini (frontier API)	.992$\pm$.006	.825$\pm$.013	—	.569$\pm$.004
F1 ↑	BAMD (ours, 8B local)	**1.000**$\pm$**.000**	**.958**$\pm$**.042**	**.860**$\pm$**.017**	**.690**$\pm$**.034**
	MC-Dropout (8B local)	1.000$\pm$.000	.903$\pm$.042	.800$\pm$.012	.622$\pm$.039
	MLE (8B local)	1.000$\pm$.000	.881$\pm$.011	.800$\pm$.005	.576$\pm$.001
	MAI-DxO (8B local)	1.000$\pm$.000	.650$\pm$.080	.690$\pm$.025	.650$\pm$.040
	ACTMED (8B local)	.958$\pm$.030	.788$\pm$.133	.501$\pm$.085	.536$\pm$.085
	LA-CDM (8B local)	.919$\pm$.021	.401$\pm$.102	.490$\pm$.125	.557$\pm$.011
	ACTMED + Nemotron-120B (local)	.913$\pm$.040	.690$\pm$.108	.717$\pm$.025	.581$\pm$.065
	ACTMED + GPT-4o (frontier API)	.998$\pm$.005	.814$\pm$.016	—	.648$\pm$.008
	ACTMED + GPT-4o-mini (frontier API)	.986$\pm$.011	.806$\pm$.015	—	.601$\pm$.002

Calibration metrics for ACTMED (120B) provide a direct comparison at a substantially larger model scale. Compared with ACTMED (8B), ACTMED (120B) improves ECE, NLL, and Brier score on heart disease, but performs worse on CKD, hepatitis, and diabetes, while BAMD remains the best-calibrated method across all four datasets. Calibration metrics for ACTMED + GPT-4o or GPT-4o-mini are not reported in (Estévez et al., 2025) and are therefore not used in this comparison. These results indicate that scaling the ACTMED surrogate alone does not systematically address the miscalibration failure mode identified in Section 1.

4.4 Adaptive budget via VOI stopping

The VOI criterion decides, per patient, whether to continue testing. Table 4 in Appendix G.2 compares fixed-budget BAMD against VOI stopping averaged over 5 seeds. VOI stopping matches fixed-budget AUC on the easier datasets and comes close to fixed-budget performance on the harder ones, while using substantially fewer tests per patient. The reduction is most evident on CKD, where the criterion terminates testing almost immediately, reflecting the saturating biomarker that already resolves the diagnosis from the prior alone.

ICL-based methods (Nori et al., 2025) maintain no probabilistic belief at all, so VOI is undefined. Among the remaining baselines, neither ACTMED nor LA-CDM is naturally compatible with VOI stopping. ACTMED's belief is implicit and recomputed per turn from token-level samples whose calibration profile is characterized in Section 4.3, yielding VOI estimates with substantially higher variance. LA-CDM's verbalized integer confidence operates at a coarser granularity than fine-grained utility comparisons require. Principled stopping thus arises naturally from BAMD's external-state architecture, whereas it requires substantial modification to retrofit onto the other paradigms.

4.5 Component ablation

To isolate the contribution of each BAMD component, we conduct four ablations (Table 5in Appendix G.2), each removing one architectural element while leaving the others intact: *w/o Fine-Tuning*

Table 2: Calibration metrics (budget B=3, over 5 seeds). Lower is better. Best in **bold**.

		CKD	Hepatitis	Heart	Diabetes
ECE ↓	BAMD (ours, 8B local)	**.026**±.002	**.108**±.010	**.062**±.006	**.065**±.008
	MC-Dropout (8B local)	.045±.012	.140±.074	.176±.016	.254±.010
	MLE (8B local)	.053±.001	.161±.027	.179±.000	.260±.004
	MAI-DxO (8B local)	.088±.015	.526±.108	.348±.038	.295±.027
	ACTMED (8B local)	.115±.009	.194±.048	.165±.019	.129±.006
	LA-CDM (8B local)	.229±.001	.272±.080	.098±.024	.127±.015
	ACTMED (120B local)	.148±.007	.249±.104	.102±.011	.230±.045
NLL ↓	BAMD (ours, 8B local)	**.028**±.004	**.204**±.073	**.391**±.029	**.545**±.002
	MC-Dropout (8B local)	.188±.072	.455±.260	.750±.060	1.06±.045
	MLE (8B local)	.210±.017	.588±.073	.776±.017	1.09±.005
	MAI-DxO (8B local)	.235±.040	2.47±.498	1.25±.111	1.03±.070
	ACTMED (8B local)	.183±.032	.553±.031	.753±.026	.639±.010
	LA-CDM (8B local)	.339±.011	.806±.169	.711±.011	.651±.012
	ACTMED (120B local)	.205±.007	.755±.251	.666±.060	.843±.045
Brier ↓	BAMD (ours, 8B local)	**.004**±.003	**.051**±.027	**.119**±.009	**.186**±.001
	MC-Dropout (8B local)	.045±.013	.110±.065	.190±.017	.261±.014
	MLE (8B local)	.051±.003	.134±.038	.196±.004	.263±.001
	MAI-DxO (8B local)	.072±.011	.526±.106	.349±.027	.287±.019
	ACTMED (8B local)	.041±.014	.184±.014	.260±.003	.204±.006
	LA-CDM (8B local)	.089±.005	.297±.075	.257±.005	.222±.003
	ACTMED (120B local)	.049±.004	.230±.078	.191±.002	.292±.021

pairs the BAMD controller with the base Llama-3.1-8B-Instruct using MC-Dropout for uncertainty; *w/o Weight Posterior (MLE)* replaces the variational posterior q_ϕ over LoRA-A with a point estimate from cross-entropy fine-tuning; *w/o External Belief* retains EIG-based selection but replaces the Bayes-rule update of Eq. equation 10 with re-querying the LLM after each outcome (implicit belief, as in ACTMED); and *w/o MI (random select)* draws tests uniformly at random while keeping the external Bayes-rule update.

The ablation results in Table 5 show that each of the four components contributes non-trivially, with the magnitude and pattern of degradation reflecting the role each plays in the framework. Removing fine-tuning collapses performance to near-random on hepatitis and diabetes, indicating that an external controller cannot compensate for an uncalibrated, task-agnostic predictive layer. Removing the weight posterior preserves accuracy on the easier datasets but degrades monotonically with difficulty, mirroring the calibration dissociation of Section 4.3: a point-estimate predictive layer ranks patients competitively under saturated likelihoods but loses ground when the likelihood remains genuinely uncertain. Removing the external belief produces the largest localized drop, on heart disease, where implicit re-querying of the LLM cannot maintain a coherent state across multiple weakly-discriminative test outcomes. Removing MI-based selection degrades AUC across all four datasets, with the gap widest on diabetes, where no single biomarker dominates the panel and informed selection therefore matters most. The four components degrade performance in distinct, non-overlapping ways, indicating that the framework's architectural commitments are complementary rather than redundant. A clinical-guideline audit of the per-step test trajectories produced by each method is provided in Appendix G.3 Table 6.

5 Conclusion

We revisited the architectural choice underlying recent LLM-based sequential clinical diagnosis, namely, placing belief and control responsibilities inside the LLM itself. BAMD reinstates the classical pairing of external Bayesian belief and information-theoretic control, and confines the LLM to supplying calibrated posterior predictive likelihoods through Bayesian fine-tuning. On four clinical benchmarks, this architectural choice produces the best-calibrated probability estimates of any LLM-based paradigm we compare against, supports principled per-patient stopping via value-of-information, and matches or exceeds prompt sampling with GPT-4o on every dataset shared with (Estévez et al., 2025), all with an 8B local model on a single consumer GPU. The limiting factor in LLM-based sequential diagnosis, we conclude, is architectural rather than scale-related.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and manuscript organization.

ETHICS STATEMENT

Our experiments use publicly available, de-identified benchmark datasets and do not involve prospective human-subject interaction. The proposed framework is intended for methodological evaluation and is not ready for clinical deployment.

REPRODUCIBILITY STATEMENT

We provide the problem formulation and method in Sections 2–3, the experimental protocol and data splits in Section 4.1, and detailed prompts, hyperparameters, algorithms, and additional experiments in Appendices C, E, F, and G.2 to support reproducibility.

REFERENCES

Hepatitis c guidance 2018 update: Aasld-idsa recommendations for testing, managing, and treating hepatitis c virus infection. *Clinical infectious diseases*, 67(10):1477–1492, 2018.

Dilip Arumugam and Thomas L Griffiths. Toward efficient exploration by large language model agents. *arXiv preprint arXiv:2504.20997*, 2025.

Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

Turgay Ayer, Oguzhan Alagoz, and Natasha K Stout. Or forum—a pomdp approach to personalize mammography screening decisions. *Operations research*, 60(5):1019–1034, 2012.

David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Nassir Navab, and Matthias Keicher. Language agents for hypothesis-driven clinical decision making with reinforcement learning. *arXiv preprint arXiv:2506.13474*, 2025.

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *Proc. of Adv. Neural Inf. Process. Syst.*, pp. 1613–1622. PMLR, 2015.

Philip Bruggmann, T Berg, ALH Øvrehus, C Moreno, Carlos Eduardo Brandão Mello, Francoise Roudot-Thoraval, RT Marinho, M Sherman, Stephen D Ryder, J Sperl, et al. Historical epidemiology of hepatitis c virus (hcv) in selected countries. *Journal of viral hepatitis*, 21:5–33, 2014.

Diabetes Care. Standards of care in diabetes—2024. *Diabetes Care*, 47:S77, 2024.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pp. 273–304, 1995.

Morris H DeGroot. *Optimal statistical decisions*. John Wiley & Sons, 2005.

Brian T Denton. Optimization of sequential decision making for chronic diseases: From data to decisions. In *Recent Advances in Optimization and Modeling of Contemporary Problems*, pp. 316–348. INFORMS, 2018.

Silas Ruhrberg Estévez, Nicolás Astorga, and Mihaela van der Schaar. Timely clinical diagnosis through active test selection. *arXiv preprint arXiv:2510.18988*, 2025.

Adam Foster, Martin Jankowiak, Elias Bingham, Paul Horsfall, Yee Whye Teh, Thomas Rainforth, and Noah Goodman. Variational bayesian optimal experimental design. *Proc. of Adv. Neural Inf. Process. Syst.*, 32, 2019.

Adam Foster, Desi R Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential bayesian experimental design. In *Proc. Intl. Conf. Mach. Learn.*, pp. 3384–3395. PMLR, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. Intl. Conf. Mach. Learn.*, pp. 1050–1059. PMLR, 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proc. Intl. Conf. Mach. Learn.*, pp. 1183–1192. PMLR, 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

KDIGO Work Group et al. Kdigo clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int Suppl*, 3(3):1, 2013.

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

James Harrison, John Willes, and Jasper Snoek. Variational bayesian last layers. *arXiv preprint arXiv:2404.11599*, 2024.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Proc. Intl. Conf. Learn. Repr.*, 1(2):3, 2022.

Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. Classification with costly features using deep reinforcement learning. In *Proc. AAAI Conf. Artif. Intell.*, volume 33, pp. 3959–3966, 2019.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Kamyar Kalantar-Zadeh, Tazeen H Jafar, Dorothea Nitsch, Brendon L Neuen, and Vlado Perkovic. Chronic kidney disease. *The lancet*, 398(10302):786–802, 2021.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew G Wilson. Large language models must be taught to know what they don't know. *Proc. of Adv. Neural Inf. Process. Syst.*, 37:85932–85972, 2024.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Proc. of Adv. Neural Inf. Process. Syst.*, 28, 2015.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Proc. of Adv. Neural Inf. Process. Syst.*, 30, 2017.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. Towards accurate differential diagnosis with large language models. *Nature*, 642(8067):451–457, 2025.

Writing Committee Members, Martha Gulati, Phillip D Levy, Debabrata Mukherjee, Ezra Amsterdam, Deepak L Bhatt, Kim K Birtcher, Ron Blankstein, Jack Boyd, Renee P Bullock-Palmer, et al. 2021 aha/acc/ase/chest/saem/scct/scmr guideline for the evaluation and diagnosis of chest pain: a report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. *Journal of the American College of Cardiology*, 78(22):e187–e285, 2021.

Abdulazeez Mousa, Waraz Mustafa, Ridwan Boya Marqas, and Shivan Hm Mohammed. A comparative study of diabetes detection using the pima indian diabetes database. *Journal of Duhok University*, 26(2):277–288, 2023.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, et al. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025.

NVIDIA. Nemotron 3 super: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning. *arXiv preprint arXiv:2604.12374*, 2026.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pp. 314–334, 2023.

Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, Laurence Aitchison, Julyan Arbel, David Dunson, Maurizio Filippone, Vincent Fortuin, Philipp Hennig, José Miguel Hernández-Lobato, et al. Position: Bayesian deep learning is needed in the age of large-scale ai. *arXiv preprint arXiv:2402.00809*, 2024.

Yu-Shao Peng, Kai-Fu Tang, Hsuan-Tien Lin, and Edward Chang. Refuel: Exploring sparse features in deep reinforcement learning for fast disease diagnosis. *Proc. of Adv. Neural Inf. Process. Syst.*, 31, 2018.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20, 2024.

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

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

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Liwen Sun, Abhineet Agarwal, Aaron Kornblith, Bin Yu, and Chenyan Xiong. Ed-copilot: Reduce emergency department wait time with language model diagnostic assistance. *arXiv preprint arXiv:2402.13448*, 2024.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*, 2024.

Abraham Wald. Sequential tests of statistical hypotheses. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 256–298. Springer, 1992.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. Blob: Bayesian low-rank adaptation by backpropagation for large language models. *Proc. of Adv. Neural Inf. Process. Syst.*, 37:67758–67794, 2024.

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*, 2018.

Deborah L Wingard, Elizabeth Barrett-Connor, et al. Heart disease and diabetes. *Diabetes in America*, 2(1):429–448, 1995.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Zhixiang Xu, Matt Kusner, Kilian Weinberger, and Minmin Chen. Cost-sensitive tree of classifiers. In *Proc. Intl. Conf. Mach. Learn.*, pp. 133–141. PMLR, 2013.

Adam X Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. *arXiv preprint arXiv:2308.13111*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *Proc. Intl. Conf. Learn. Repr.*, 2022.

Supplementary File

A Related work

Sequential clinical test selection has been studied long before the LLM era. The cost-sensitive feature acquisition literature formulates it as reinforcement learning (Peng et al., 2018; Janisch et al., 2019), POMDP planning (Ayer et al., 2012), or tree-structured policies (Xu et al., 2013) over tabular classifiers, with explicit belief state and cost-aware objective. These systems establish the principled machinery of Bayesian decision theory for sequential diagnosis on structured features, and a natural question is how to carry this machinery into settings where unstructured clinical text and broad prior medical knowledge, which modern LLMs encode, also matter.

LLM-based methods extend sequential test selection to natural language along several directions, each effective in its intended goal. Surrogate-based Bayesian experimental design, exemplified by ACTMED (Estévez et al., 2025), demonstrates that a frontier LLM can serve as a zero-shot conditional outcome sampler, selecting tests via expected KL divergence without task-specific training; clinician-in-the-loop evaluations confirm its clinical plausibility. Reinforcement-learning-based test ordering, exemplified by LA-CDM (Bani-Harouni et al., 2025), trains an LLM agent to emit disease predictions alongside integer confidence scores and to select tests via prompted deliberation (Yao et al., 2022), yielding an interpretable verbalized interface suited to coarse risk stratification. Prompted multi-agent orchestration, exemplified by MAI-DxO (Nori et al., 2025), achieves strong accuracy on challenging NEJM case reports through structured deliberation among off-the-shelf frontier models, demonstrating the power of frontier-model reasoning on complex differential diagnosis. Further work spans emergency-department triage (Sun et al., 2024) and sequential policy optimization on tabular clinical data (Denton, 2018). Despite the diversity of techniques, these paradigms share a common architectural pattern: belief maintenance and test-selection decisions reside inside the LLM, whether as context accumulation, verbalized confidence, or agent deliberation. In single-turn clinical prediction, this pattern has produced strong results, but in sequential diagnosis, it exposes the system to two failure modes that the broader LLM literature has characterized as persistent across scale. The first is miscalibration: LLM probability estimates in clinical settings systematically diverge from empirical frequencies, and prompting-based elicitation reduces but does not eliminate the gap (Savage et al., 2024; Xiong et al., 2023; Kapoor et al., 2024). The second is hallucination: models generate confidently incorrect clinical content whose provenance cannot be audited from the model's own outputs alone (Hager et al., 2024; Pal et al., 2023). In a single-turn setting, either failure mode affects only a local output, but in sequential diagnosis, both become compounding: the current step's probability enters the next step's belief update, and the current step's reasoning enters the next step's context. BAMD's contribution is to show that relocating belief and decision outside the LLM is sufficient to prevent both from propagating across turns.

LLM calibration and uncertainty quantification are broadly relevant because any LLM-based diagnostic system consumes the LLM's uncertainty estimates, implicitly or explicitly. LLM predictions in clinical settings exhibit well-documented overconfidence (Savage et al., 2024; Papamarkou et al., 2024; Xiong et al., 2023), which prompting-based elicitation can reduce but not eliminate (Kapoor et al., 2024). Among fine-tuning approaches that retain a posterior over parameters, variational inference over LoRA adapters (Wang et al., 2024), post-hoc Laplace approximations (Yang et al., 2023), variational Bayesian last layers (Harrison et al., 2024), MC dropout (Gal & Ghahramani, 2016), and deep ensembles (Lakshminarayanan et al., 2017) all yield predictive distributions from which calibrated probabilities can be derived. We instantiate the predictive layer of BAMD using BLoB (Wang et al., 2024); the framework is agnostic to this choice, and any method that yields a posterior over adapter parameters is compatible. What matters for sequential diagnosis is not which Bayesian method is used, but that the calibrated signals it produces are consumed by an external Bayesian controller rather than by the LLM itself.

Bayesian experimental design provides the control-layer machinery that BAMD reinstates. Test selection by expected information gain originates in the classical BED literature (Lindley, 1956; Chaloner & Verdinelli, 1995), and the same mutual information criterion serves as the acquisition function in Bayesian optimization (Srinivas et al., 2009) and active learning (Houlsby et al., 2011; Gal et al., 2017). Scalable approximations via variational bounds (Foster et al., 2019) and amortized policy networks (Foster et al., 2021) extend BED to higher-dimensional settings. BAMD and ACTMED

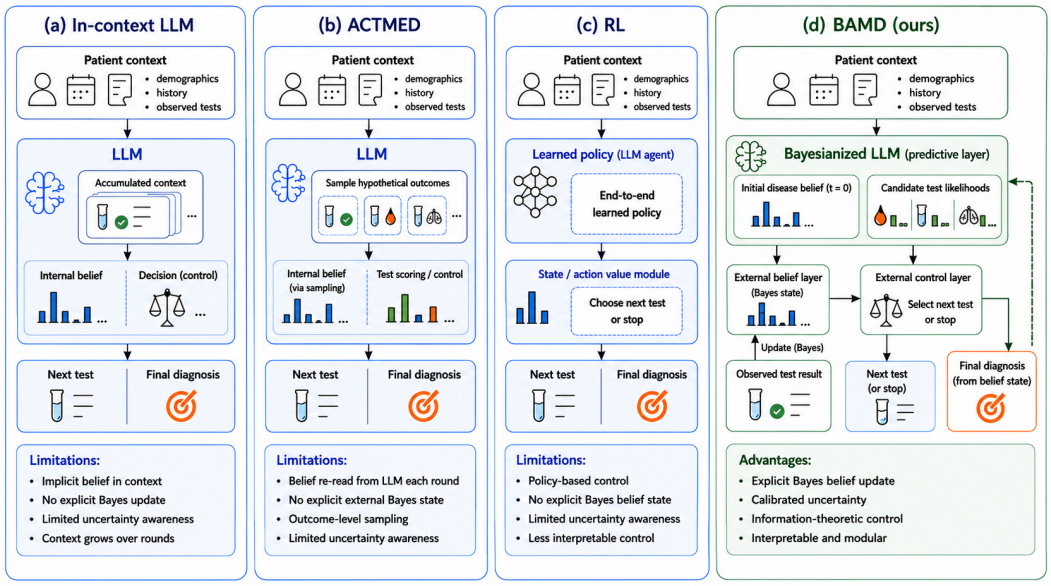


Figure 3: Architectural comparison of representative paradigms for LLM-based sequential diagnosis. (a) In-context learning keeps belief and control implicitly inside the accumulating context of the LLM. (b) ACTMED uses the LLM as a conditional outcome sampler, but still lacks an explicit external Bayesian belief state. (c) RL-based agents learn a policy for test selection and stopping, without maintaining an explicit Bayes state. (d) BAMD separates prediction from decision-making: a Bayesianized LLM provides calibrated predictive signals, while an external Bayesian belief layer performs explicit Bayes updates and an external control layer selects the next test or stops.

can be seen as two instantiations of BED with LLMs that differ in the source of stochasticity and the locus of belief. ACTMED samples over-generated token outcomes, which are well-suited to zero-shot frontier-model deployment where no weight-space posterior is accessible; BAMD samples over the model's *weight* posterior, producing likelihood estimates whose calibration is regularized by the variational objective and consumed by an explicit external Bayes-rule update. The two approaches target different deployment regimes, namely no-training frontier APIs versus fine-tunable local models, and the architectural choices in each reflect that difference.

Separating prediction from control in LLM agents has emerged as a broader theme beyond clinical diagnosis. The clearest precedent for our architectural move is (Arumugam & Griffiths, 2025), who show that using LLMs to explicitly implement the subroutines of posterior sampling for RL (sampling, updating, optimal-policy computation) yields substantially better exploration than designs that rely on the LLM to implicitly induce an exploration strategy. The useful distinction, which we adopt, is between an LLM that *implements* a classical algorithm and an algorithm that happens to be *implemented with* an LLM. BAMD sits firmly in the latter category: the belief is an external Bayesian object, test selection is a classical BED operator, and the LLM is a calibrated likelihood source.

B LLM-BASED APPROACHES FOR SEQUENTIAL DIAGNOSIS

Here we review two LLM-based paradigms for sequential diagnosis that inform the design of BAMD and serve as primary baselines in our experiments. Both define concrete policies for choosing the next test index $j^*_{n,t}$ and producing the final diagnosis $\hat{d}_n$ in Eq. equation 2.

The first paradigm, in-context learning (ICL), treats the LLM as a black-box policy. At each turn t, the LLM receives the patient vignette $(\mathbf{x}_n,\ \mathcal{O}_{n,t-1})$ and the list of remaining test names indexed by $\mathcal{J}_{n,t-1} := \{1,\dots,M\} \setminus \mathcal{S}_{n,t-1}$, and is asked to name the most informative next test. Letting π_{ICL} denote the resulting LLM-induced categorical distribution over test indices, the ICL policy sets (Nori

et al., 2025)

$$j^*_{n,t} \sim \pi_{\text{ICL}}\Big(\cdot \,\big|\, \mathbf{x}_n,\, \mathcal{O}_{n,t-1};\, \hat{\boldsymbol{\theta}}\Big), \qquad \hat{d}_n = \arg\max_{d\in\mathcal{D}} p\Big(D_n = d \,\big|\, \mathbf{x}_n,\, \mathcal{O}_{n,\tau_n};\, \hat{\boldsymbol{\theta}}\Big), \tag{14}$$

where $\hat{\boldsymbol{\theta}}$ denotes the (fixed) LLM parameters and the final risk is read from the model's logits on a separate Yes/No prompt. Prediction, belief maintenance, and test selection are entangled inside the LLM: the belief state exists implicitly in the context window, without an external representation that can be updated by Bayes rule or coupled with a formal stopping rule.

The second paradigm, proposed by (Estévez et al., 2025) and referred to here as ACTMED, retains the classical Bayesian experimental design (BED) objective but uses the LLM as a conditional outcome sampler rather than as a fitted likelihood model. For each candidate test index $j \in \mathcal{J}_{n,t-1}$ at turn t, the LLM is queried N_{mc} times in generation mode with temperature $\gamma > 0$ to produce a hypothetical outcome

$$u^{(m)}_{n,j} \sim p^{\text{LLM}}\Big(\cdot \,\big|\, \mathbf{x}_n,\, \mathcal{O}_{n,t-1},\, T_j;\, \hat{\boldsymbol{\theta}},\, \gamma\Big), \qquad m = 1, \ldots, N_{\text{mc}}. \tag{15}$$

For each sampled outcome, a second-generation mode query returns a scalar risk estimate

$$p^{(m)}_{\text{post},\,j} \approx p\Big(D_n = 1 \,\big|\, \mathbf{x}_n,\, \mathcal{O}_{n,t-1},\, u^{(m)}_{n,j};\, \hat{\boldsymbol{\theta}}\Big), \qquad m = 1, \ldots, N_{\text{mc}}, \tag{16}$$

and, with p_{prior} denoting the prior risk obtained by querying the LLM on the vignette $(\mathbf{x}_n,\, \mathcal{O}_{n,t-1})$ alone, each candidate test is scored by the empirical expected KL divergence between posterior and prior Bernoulli beliefs,

$$\text{EIG}^{\text{ACTMED}}_{n,t}(j) = \frac{1}{N_{\text{mc}}} \sum_{m=1}^{N_{\text{mc}}} \text{KL}\Big(\mathbb{B}(p^{(m)}_{\text{post},\,j}) \,\big\|\, \mathbb{B}(p_{\text{prior}})\Big), \tag{17}$$

with the selected test given by $j^*_{n,t} = \arg\max_{j\in\mathcal{J}_{n,t-1}} \text{EIG}^{\text{ACTMED}}_{n,t}(j)$. This scheme departs from our framework in two concrete ways that motivate the design of BAMD. The prior p_{prior} is recomputed at every turn by re-querying the LLM on the updated vignette, so belief updating proceeds implicitly through context accumulation rather than through an explicit Bayes-rule operator on an external state. The stochasticity used to estimate the expectation in Eq. equation 17 arises entirely from token-level temperature sampling, and collapses to a degenerate estimator when the surrogate is a fine-tuned classifier with sharply peaked logits or when temperature is small.

C PROMPT DESIGN

The LLM backbone is queried through two prompt types, each serving a distinct role in the BAMD pipeline. We first describe the prompt types, then explain how they are used during fine-tuning and during inference.

Prompt Type A (likelihood estimation). Given a patient vignette $(\mathbf{x}, \mathcal{O}_{t-1})$ and a disease hypothesis d, this prompt asks whether a candidate test T_j would produce an abnormal result. The model output provides the class-conditional likelihood $p^{\text{LLM}}(Z_j{=}\text{abn} \mid d,\, \mathbf{x},\, \mathcal{O}_{t-1})$ required by Eq. equation 6. An example prompt for $d{=}1$ reads: "This patient has [disease]. [vignette]. Would [test] be abnormal? Yes or No." A corresponding prompt for $d{=}0$ replaces the condition with "does NOT have [disease]."

Prompt Type B (risk prediction). Given a patient vignette $(\mathbf{x}, \mathcal{O}_{t-1})$ without conditioning on D, this prompt asks whether the patient has the disease. The model output provides the risk estimate $p(D{=}1 \mid \mathbf{x},\, \mathcal{O}_{t-1})$ required by Eq. equation 5. An example reads: "[vignette]. Does this patient have [disease]? Yes or No."

Fine-tuning. Both BLoB and the MLE baseline are fine-tuned on the same training data constructed from patients indexed by $n = 1, \ldots, N_{\text{train}}$. For each training patient, we generate three sample types using ground-truth disease status and clinical reference ranges as labels: (i) Type A with $D{=}1$, (ii) Type A with $D{=}0$, and (iii) Type B. To expose the model to partial observations, we vary the subset of revealed tests per patient, yielding approximately 4 training samples per patient. BLoB fine-tuning optimizes the ELBO in Eq. equation 3; MLE fine-tuning optimizes the standard cross-entropy loss.

Inference. At test time, BAMD and ACTMED use the prompts differently. BAMD issues one Type B query at turn $t{=}0$ to initialize the belief (Eq. equation 5), then at each subsequent turn issues $2|\mathcal{J}_{t-1}|$ Type A queries (one per candidate test per disease hypothesis) to compute the likelihoods in Eq. equation 6 for EIG-based test selection (Eq. equation 8). All subsequent belief updates follow Eq. equation 10 without re-querying the LLM for risk prediction. ACTMED, by contrast, re-issues Type B queries at every round to obtain updated risk estimates, and uses outcome-level Monte Carlo sampling rather than conditioning on D.

Clinical vignette generation. Patient features are rendered as structured natural-language summaries (e.g., "A 45-year-old male patient presents with blood pressure 140/90 mmHg. Serum creatinine was measured at 1.2 mg/dL."). At each turn, newly revealed test results are appended to the vignette. All methods observe the same ground-truth values and build identical vignettes.

D TRAINING DATA

Let $\{(\mathbf{x}_n,\, D_n,\, \{z_{n,j}\}_{j=1}^{M})\}_{n=1}^{N_{\text{tr}}}$ denote a set of N_{tr} fully-observed training patients with known disease status D_n and known test outcomes $z_{n,j}$ for every test $T_j \in \mathcal{T}$. From these patients, we construct a supervised prompt-answer dataset

$$\mathcal{D}_{\text{tr}} := \{(\boldsymbol{\varphi}_i,\, y_i)\}_{i=1}^{N_q}, \qquad \boldsymbol{\varphi}_i \in \Sigma^*,\; y_i \in \{\text{Yes}, \text{No}\}, \tag{18}$$

in which Σ denotes the natural-language token space, each prompt $\boldsymbol{\varphi}_i$ renders a (partial) patient vignette together with a Yes/No question, and each answer y_i is derived from ground-truth clinical reference ranges and the known disease label D_n. We use two prompt types in $\mathcal{D}_{\text{tr}}$, corresponding to the two predictive quantities the downstream controller in Section 3 consumes: disease-conditional test-outcome likelihoods and disease risk estimates.

The first type, which we refer to as *Type A*, conditions on a disease hypothesis $D_n = d$ and a candidate test T_j, and asks whether the test would be abnormal. A disease-positive instance from the chronic kidney disease (CKD) training set reads

> *"This patient has chronic kidney disease. The patient is 62 years old. The patient has hypertension. The patient does not have diabetes mellitus. Would their serum creatinine result be abnormal? Yes or No."*

with ground-truth answer $y_i = \text{Yes}$ when the true creatinine value falls outside the clinical reference range and No otherwise. A disease-negative prompt replaces the opening sentence with "This patient does NOT have chronic kidney disease." and reuses the same vignette and answer derivation. Type A prompts supply the class-conditional likelihood $p(Z_j = z \mid D_n = d,\, \mathbf{x}_n,\, \mathcal{O}_{n,t-1})$ that the test-selection controller in Section 3.3 queries.

The second type, *Type B*, omits the disease hypothesis and asks directly whether the patient has the disease. An example from the same training set reads

> *"The patient is 62 years old. The patient has hypertension. The patient does not have diabetes mellitus. Does this patient have chronic kidney disease? Yes or No."*

with ground-truth answer determined by the disease label D_n. Type B prompts supply the initial risk $p(D_n = d \mid \mathbf{x}_n)$ used by the belief initialization in Section 3.2.

Across both types, patient features are rendered as structured natural-language summaries. We always include demographic fields such as age, sex, and presenting signs in the vignette, and append revealed test results, when present, as sentences of the form "[test name] was measured at [value] [unit]." For each training patient, we generate approximately four prompt-answer pairs by varying the subset of revealed tests across samples, so that the model learns to predict an unseen test outcome from partial evidence, which is the exact setting encountered at inference time. This yields a total of $N_q \approx 4\, N_{\text{tr}}$ samples. The full prompt catalog across all four datasets is given in Appendix C.

Table 3: Hyperparameters for LLM-based methods.

Method	Configuration	Value
BAMD (BLoB)	LoRA rank / α	$r = 8$, $\alpha = 16$
	Target modules	`q_proj`, `v_proj`, `lm_head`
	Prior scale / KL lr	$\beta = 0.2$, $\eta_{\text{KL}} = 0.02$
	Optimizer	AdamW, warmup 0.06, batch 8
MLE	LoRA rank / α	$r = 8$, $\alpha = 16$
	Target modules	same as BAMD
	Optimizer	AdamW, warmup 0.06, batch 8
ACTMED (8B)	Surrogate checkpoint	Base LlaMA
	Sampling temp. / #samples	$\gamma = 1.0$, $N_{\text{mc}} = 10$
LA-CDM	LoRA rank / α	$r = 8$, $\alpha = 16$, dropout 0.05
	Target modules	all attention + MLP projections
	GRPO config	lr 10^{-5}, $G = 4$, $\beta_{\text{KL}} = 0.04$, 50 steps/obj/epoch

E EXPERIMENT IMPLEMENT DETAILS

F BAMD ALGORITHM

Algorithm 1 summarizes the variational fine-tuning procedure that produces the predictive layer of BAMD. The base model $\mathbf{W}_0$ is frozen throughout; only the variational parameters $\boldsymbol{\phi} = \{\boldsymbol{\mu}_\ell, \boldsymbol{\rho}_\ell, \mathbf{B}_\ell\}_{\ell=1}^{L}$ of the LoRA adapters are updated. At each minibatch, a single Monte Carlo sample is drawn from $q_{\boldsymbol{\phi}}$ via local reparameterization with flipout (Kingma et al., 2015; Wen et al., 2018), and the negative ELBO of Eq. equation 3 is minimized with the KL term reweighted by π_t following (Blundell et al., 2015). The output of the procedure is the trained variational posterior $q_{\boldsymbol{\phi}^*}$, from which posterior weight samples are drawn at inference time and consumed by Algorithm 2.

Algorithm 2 summarizes the full BAMD pipeline for a generic patient, with the patient index n suppressed per the convention of Section 2. Posterior weight samples $\{\boldsymbol{A}^{(k)}\}_{k=1}^{K}$ are drawn once from the variational posterior $q_{\boldsymbol{\phi}}$ at the start of the episode and reused throughout the loop, both for the belief initialization (line 3) and for every per-turn likelihood estimate (line 7), so the per-patient cost of Bayesian inference scales linearly with K rather than with the budget B or the number of tests M. Each turn computes the same set of class-conditional likelihoods $\hat{p}_j(z \mid d)$ once, then derives the marginal predictive, the expected information gain, and the value of information from these shared quantities. The VOI stopping rule (line 12) is checked before test selection (line 14), so the agent can terminate without ordering a further test whenever no remaining candidate exceeds its acquisition cost.

Algorithm 1 Bayesian fine-tuning of LoRA adapters via variational inference

Require: Frozen base model $\mathbf{W}_0$; supervised dataset $\mathcal{D}_{\text{tr}}$ defined in Eq. equation 18; LoRA target layers $\ell = 1, \ldots, L$; prior scale β; learning rate η; number of epochs E; minibatch size b
Ensure: Trained variational posterior $q_{\boldsymbol{\phi}^*}$
1: Initialize $\boldsymbol{\mu}_\ell$ via standard LoRA initialization, $\boldsymbol{\rho}_\ell$ to a small value, and $\mathbf{B}_\ell \leftarrow \mathbf{0}$ for $\ell = 1, \ldots, L$
2: **for** epoch $= 1, \ldots, E$ **do**
3: **for** minibatch $\mathcal{B} \subset \mathcal{D}_{\text{tr}}$ of size b **do**
4: Compute KL reweighting coefficient π_t
5: Sample $\mathbf{A}_\ell \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \operatorname{diag}(\boldsymbol{\sigma}_\ell^2))$ with $\boldsymbol{\sigma}_\ell = (\boldsymbol{\rho}_\ell)^2$, for $\ell = 1, \ldots, L$
6: Compute $\mathcal{L}_{\text{ELBO}}(\boldsymbol{\phi})$ on $\mathcal{B}$ via Eq. equation 3
7: Update $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \eta \nabla_{\boldsymbol{\phi}} \mathcal{L}_{\text{ELBO}}(\boldsymbol{\phi})$
8: **end for**
9: **end for**
10: **return** $q_{\boldsymbol{\phi}^*}$

Algorithm 2 Bayesian Agentic Medical Diagnosis (BAMD) for a generic patient

Require: Patient features $\mathbf{x}$; test panel $\{T_j\}_{j=1}^{M}$ with costs $\{c_j\}_{j=1}^{M}$; variational posterior $q_{\boldsymbol{\phi}}$ over LoRA parameters; number of posterior samples K; budget B; action set $\mathcal{A}$ and utility $u : \mathcal{A} \times \mathcal{D} \to \mathbb{R}$

Ensure: Final diagnosis $\hat{d} \in \mathcal{D}$

Draw $\tilde{\mathbf{A}}^{(k)} \sim q_{\boldsymbol{\phi}}$ for $k = 1, \ldots, K$
Initialize $\mathcal{S} \leftarrow \emptyset$, $\mathcal{O} \leftarrow \emptyset$
Compute initial belief $\hat{r}(d)$ for $d \in \mathcal{D}$ via Eq. equation 5
while $|\mathcal{S}| < B$ **do**
  $\mathcal{J} \leftarrow \{1, \ldots, M\} \setminus \mathcal{S}$
  for $j \in \mathcal{J}$ **do**
    Compute class-conditional likelihoods $\hat{p}_j(z \mid d)$ via Eq. equation 6
    Compute marginal predictive $\hat{p}_j^{\text{mar}}(z)$ via Eq. equation 7
    Compute $\text{EIG}(j)$ via Eq. equation 8
    Compute $\text{VOI}(j)$ via Eq. equation 12
  end for
  if $\max_{j \in \mathcal{J}} [\text{VOI}(j) - c_j] < 0$ **then**
    break VOI stopping, Eq. equation 13
  end if
  Select next test j^* via Eq. equation 9
  Apply test T_{j^*} and observe outcome $z^* \in \mathcal{Z}$
  Update belief $\hat{r}(d)$ for $d \in \mathcal{D}$ via Eq. equation 10
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{j^*\}, \quad \mathcal{O} \leftarrow \mathcal{O} \cup \{(j^*, z^*)\}$
end while
return $\hat{d} \leftarrow \arg\max_{d \in \mathcal{D}} \hat{r}(d)$ Eq. equation 2

Table 4: BAMD with VOI stopping vs. fixed budget B=3.

		CKD	Hepatitis	Heart	Diabetes
AUC ↑	Fixed B=3	1.000±.000	.975±.025	.907±.011	.800±.025
	VOI stopping	1.000±.000	.948±.074	.892±.022	.770±.013
Avg tests ↓	Fixed B=3	3.00±.00	3.00±.00	3.00±.00	3.00±.00
	VOI stopping	**0.23**±.04	**2.18**±.06	**2.62**±.07	**2.54**±.04

G Additional Experiments

G.1 Dataset details

Chronic Kidney Disease (CKD) (Kalantar-Zadeh et al., 2021) contains 157 patients with age, blood pressure, and specific gravity as known features, and seven candidate diagnostic tests. Serum creatinine serves as a near-deterministic biomarker, making this the easiest dataset.

Hepatitis C (Bruggmann et al., 2014) comprises 112 patients (56 HCV-positive, 56 healthy) with age and sex as known features, along with ten liver-function tests (ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, PROT). The class balance matches the cohort used by (Estévez et al., 2025), enabling an identical-cohort comparison.

Heart Disease (Cleveland) (Wingard et al., 1995) consists of 746 patients with age and sex as known features, along with nine tests: chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, ST depression, and ST slope.

Diabetes (Pima Indians) (Mousa et al., 2023) contains 768 patients with age and number of pregnancies as known features, and six diagnostic tests. The class imbalance (35% disease prevalence) combined with the absence of a single dominant biomarker makes this the most challenging dataset.

Table 5: Component ablation (budget B=3, mean AUC over 5 seeds).

	CKD	Hepatitis	Heart	Diabetes
BAMD (full)	**1.000**	**0.975**	**0.907**	**0.800**
w/o Fine-Tuning	0.911	0.407	0.785	0.514
w/o Weight Posterior (MLE)	1.000	0.896	0.822	0.693
w/o External Belief	0.983	0.898	0.595	0.730
w/o MI (random select)	0.943	0.851	0.770	0.667

Table 6: Full 3-step test selection trajectory (most frequent test at each step, % of patients).

		CKD	Hepatitis	Heart	Diabetes
1st test	BAMD	Creatinine (88%)	AST (100%)	ST Slope (91%)	Glucose (97%)
	MLE	Creatinine (75%)	CHOL (43%)	ChestPain (34%)	Glucose (41%)
	MAI-DxO	bgr (93%)	ALT (100%)	RestingECG (89%)	DiabPedigree (45%)
	ACTMED	Creatinine (66%)	AST (48%)	ChestPain (81%)	Pregnancies (49%)
	LA-CDM	bp (100%)	ALT (48%)	ChestPain (100%)	Glucose (99%)
2nd test	BAMD	Blood Urea (50%)	GGT (86%)	ChestPain (59%)	BMI (60%)
	MLE	Hemoglobin (59%)	GGT (47%)	ST Slope (47%)	Glucose (35%)
	MAI-DxO	Hemoglobin (36%)	GGT (100%)	ExerAngina (64%)	Glucose (42%)
	ACTMED	bu/pot (25%/25%)	CREA (82%)	RestingECG (24%)	DiabPedigree (42%)
	LA-CDM	bgr (100%)	ALP (39%)	RestingBP (100%)	BloodPressure (97%)
3rd test	BAMD	Sodium (37%)	CHE (52%)	ExerAngina (46%)	DiabPedigree (31%)
	MLE	Sodium (62%)	CREA (43%)	ExerAngina (36%)	BMI (30%)
	MAI-DxO	Hemoglobin (26%)	AST (100%)	Cholesterol (54%)	BMI (31%)
	ACTMED	Hemoglobin (21%)	PROT (21%)	FastingBS (30%)	SkinThick (22%)
	LA-CDM	bu (100%)	ALT (43%)	Cholesterol (100%)	SkinThick (97%)

G.2 Additional Analysis of Stopping and Components

G.3 Clinical validity of test selection

Table 6 audits per-step modal test selections, averaged over 5 seeds, against the most informative panel-available tests recommended by the corresponding clinical guidelines: serum creatinine for CKD (Group et al., 2013), hepatic enzymes such as AST/ALT for hepatitis C monitoring (aas, 2018), exercise-induced ECG features such as ST slope for stable ischemic heart disease (Members et al., 2021), and plasma glucose for diabetes (Care, 2024). BAMD's modal first test coincides with these recommendations on every dataset, and its subsequent steps form clinically coherent panels (renal function, liver enzyme, cardiac risk, metabolic) that a guideline-following practitioner would recognize.

The baselines deviate from guideline-consistent selection in ways that reflect their architectural choices. MLE, despite sharing BAMD's external EIG controller, selects cholesterol as the hepatitis first-line test with markedly lower per-patient consistency, isolating the effect of likelihood calibration on selection quality. MAI-DxO's prompted multi-agent deliberation chooses non-guideline tests on CKD (blood glucose) and diabetes (diabetes pedigree), reflecting prompt-level associations rather than disease-conditional information. The 8B ACTMED selects pregnancies, a risk factor rather than a diagnostic test, as the first test for diabetes for nearly half the cohort, and prioritizes creatinine over hepatic enzymes as the second test for most hepatitis patients. LA-CDM emits a near-identical template across patients within each dataset, exhibiting little per-patient adaptation. We do not claim equivalence with a practicing clinician; guideline adherence is citable and reproducible, and complements the subjective clinician-in-the-loop evaluation of (Estévez et al., 2025).

G.4 Computational cost

We characterize each method by the number of LLM queries it issues per patient and the mode of those queries. At each decision turn t, BAMD issues $2 \cdot |\mathcal{J}_{t-1}| \cdot K$ Type A forward passes per Eq. equation 6, plus K Type B passes for the belief initialization at $t = 1$. Each pass reads the

Table 7: Qwen 2.5-7B-Instruct results (budget B=3). Best in **bold**.

		CKD	Hepatitis	Heart	Diabetes
AUC ↑	BAMD (ours)	1.000±0.000	0.967±0.015	0.919±0.025	0.786±0.007
	MLE	1.000±0.000	0.925±0.025	0.823±0.012	0.716±0.013
	MC-Dropout	1.000±0.000	0.935±0.017	0.812±0.003	0.746±0.011
	MAI-DxO	0.998±0.003	0.758±0.021	0.783±0.015	0.712±0.016
	ACTMED	0.865±0.019	0.764±0.009	0.567±0.004	0.689±0.030
	LA-CDM	0.969±0.031	0.617±0.025	0.518±0.016	0.661±0.003
ECE ↓	BAMD (ours)	0.018±0.003	0.097±0.005	0.062±0.004	0.061±0.001
	MLE	0.052±0.002	0.177±0.012	0.181±0.017	0.153±0.002
	MC-Dropout	0.041±0.004	0.152±0.014	0.110±0.029	0.145±0.002
	MAI-DxO	0.061±0.001	0.512±0.131	0.229±0.030	0.267±0.019
	ACTMED	0.183±0.005	0.248±0.059	0.215±0.004	0.152±0.009
	LA-CDM	0.256±0.002	0.359±0.120	0.171±0.014	0.152±0.001

Table 8: Computational cost at B=3 on a single consumer GPU (RTX 6000 Ada). Training time includes fine-tuning only (ACTMED requires none). Inference time is wall-clock for the full test set. $K = 10$ posterior samples for BAMD, $N_{\text{mc}} = 10$ outcome samples for ACTMED.

		Calls/patient	Mode	CKD	Hep.	Heart	Diab.
Train	BAMD (ours)	—	—	7 min	5 min	15 min	10 min
	MLE	—	—	12 min	12 min	75 min	55 min
	LA-CDM	—	—	45 min	45 min	300 min	240 min
Inference	BAMD (ours)	$\mathcal{O}(B\|\mathcal{T}\|K)$	logit read	26 min	28 min	202 min	100 min
	MLE	$\mathcal{O}(B\|\mathcal{T}\|)$	logit read	2 min	2 min	136 min	7 min
	MC-Dropout	$\mathcal{O}(B\|\mathcal{T}\|K)$	logit read	25 min	25 min	200 min	90 min
	ACTMED (8B)	$\mathcal{O}(B\|\mathcal{T}\|N_{\text{mc}})$	generation	71 min	85 min	386 min	266 min
	ACTMED (120B)	$\mathcal{O}(B\|\mathcal{T}\|N_{\text{mc}})$	generation	200 min	225 min	575 min	887 min
	LA-CDM	$\mathcal{O}(B)$	generation	11 min	5 min	28 min	40 min

Yes-token logit at a single output position rather than generating tokens autoregressively. The total per-patient cost at budget B is

$$K + \sum_{t=1}^{B} 2\,|\mathcal{J}_{t-1}|\,K \;=\; \mathcal{O}(B \cdot |\mathcal{T}| \cdot K), \tag{19}$$

which mirrors the per-patient complexity reported by (Estévez et al., 2025) for ACTMED, with their MC sample count N_{mc} replaced by the BAMD posterior sample count K.

Table 8 reports per-patient call complexity together with wall-clock evaluation time across the four LLM-based methods. The runtimes reflect two structural distinctions in the call counts. BAMD and MLE issue single-token logit reads, while ACTMED and LA-CDM run the LLM in autoregressive generation mode, which produces the runtime ordering between the two groups at matched nominal sample counts. ACTMED (120B) further increases wall-clock inference time over ACTMED (8B), while still trailing the 8B BAMD model in diagnostic performance, showing that scaling the surrogate alone does not close the performance gap. Within each group, BAMD's K-fold overhead over MLE corresponds to marginalizing over the variational posterior in Eq. equation 4, which is the same operation that yields the calibration improvements reported in Table 2. The runtime is dominated by per-call cost rather than by the size of the test panel, so absolute differences scale with cohort size rather than with $|\mathcal{T}|$.

The runtimes in Table 8 describe inference on local hardware, where compute is a one-time hardware cost. For methods accessed through commercial APIs, the corresponding axis is recurring per-query expenditure. As an illustrative reference, Table 9 reports the estimated API cost of running ACTMED + GPT-4o under our experimental protocol across the four test cohorts in Table 1. At current OpenAI pricing of $2.50 per million input tokens and $10.00 per million output tokens, and at an average

[2] `https://openai.com/api/pricing/`, accessed April 2026.

Table 9: Estimated API cost for ACTMED + GPT-4o under our experimental protocol ($80/20$ split, 5 seeds, $B = 3$, $N_{\text{mc}} = 10$). Per-call token assumption: ≈ 450 input $+ \approx 80$ output. OpenAI GPT-4o pricing: \$2.50 per million input tokens, \$10.00 per million output tokens.[2]

Dataset	$\lvert\mathcal{T}\rvert$	**Calls/patient**	**Total calls**	**Cost**
CKD	7	363	58,080	\$112
Hepatitis	10	543	59,730	\$115
Heart	9	483	359,835	\$693
Diabetes	6	303	233,310	\$449
Total			**710,955**	**\$1,369**

prompt length of ≈ 450 input and ≈ 80 output tokens per call, the four-dataset total is approximately \$1,369 per full evaluation pass, with the per-dataset breakdown dominated by cohort size. The cost is incurred in full each time the evaluation is rerun, so every ablation in Section 4.5, hyperparameter sweep, or seed extension repeats the same expenditure. The BAMD inference stack runs entirely on a single consumer GPU with no recurring per-query cost and no transmission of patient vignettes off premises, which is a practical prerequisite in clinical deployment settings where data residency or regulatory constraints limit the use of commercial APIs.