Adaptive Bayesian Active Querying with LLMs for Efficient Information Gathering

Ognjen Malkoc
Gaudiy Inc.
ognjen.malkoc@gaudiy.com

Hongtao Hao
University of Wisconsin Madison
hongtaoh@cs.wisc.edu

Grisha Szep
Chiba Institute of Technology, Japan
grisha.szep@chibatech.ac.jp

Shubham Saha
Gaudiy Inc.
shubham.saha@gaudiy.com

Mizuki Oka
Gaudiy Inc.
Chiba Institute of Technology, Japan
oka.mizuki@chibatech.ac.jp

Joseph Austerweil
Chiba Institute of Technology, Japan
joseph.austerweil@chibatech.ac.jp

Abstract

We propose a hybrid framework for cost-aware sequential diagnosis that combines LLM-generated questions with principled Bayesian belief tracking and decision-making. LLMs can generate plausible queries and estimate likelihoods, but lack explicit, calibrated beliefs—leading to inefficient information gathering and poorly grounded decisions about when to order costly tests. Our hybrid approach uses an LLM to propose candidate questions and estimate conditional probabilities, while a Bayesian decision layer maintains beliefs and selects queries by expected information gain per unit cost. This separation enables cost-sensitive interviewing: low-cost questions narrow the hypothesis space before high-cost tests are recommended, and tests are ordered only when sufficient probability mass concentrates on hypotheses they would confirm or rule out. To handle cases where standard information gain stalls due to redundant questions from the LLM, we introduce an adaptive mechanism that switches to an information metric that focuses on the top candidates. On a medical diagnosis benchmark using locally deployable LLMs, our framework achieves high diagnostic accuracy while reducing unnecessary queries, demonstrating that principled uncertainty tracking enables robust, cost-aware diagnosis. Reproducible code is available at https://github.com/gaudiy/rnd_bayesian_eig_colm.

1 Introduction

A clinician must diagnose an emergency patient under tight time and cost constraints. Which question should they ask first, and when should they stop? This is one instance of a general problem: an agent must gather information through a sequence of questions, an interview, whose answers are uncertain and whose costs vary. Bayesian experimental design offers a principled answer (Rainforth et al., 2024; Cover, 1999): by maintaining a posterior over hypotheses, we estimate every question's expected information gain and ask the one expected to reduce uncertainty most.

Implementing this strategy, however, requires a calibrated likelihood model $P(a \mid q, h)$ for each answer a, question q, and hypothesis h. In closed domains these likelihoods can be precomputed; in open domains, the relevant questions are not known in advance and constructing calibrated tables quickly becomes prohibitive. Large language models (LLMs) offer a way around this bottleneck: trained on diverse data, they implicitly encode broad

world knowledge (Li et al., 2023; Gurnee & Tegmark, 2024; Huang et al., 2022) and can generate plausible questions and likelihood estimates from natural-language context.

Building on these capabilities, recent work extends Bayesian experimental design by using an LLM to generate candidate questions and estimate likelihoods on demand, enabling expected information gain (EIG) estimation without calibrated lookup tables (Choudhury et al., 2025). This complements broader work on LLMs with probabilistic programs and world-model interfaces (Lew et al., 2023; Wong et al., 2023; Zhao et al., 2024), and enables Bayesian interviews beyond closed domains.

However, using LLMs for both question generation and likelihood estimation introduces new failure modes. Although EIG provides a principled criterion for selecting informative questions, LLM-generated candidates and likelihood estimates may not serve the information-gathering objective. The candidates can be redundant or insufficiently discriminative, especially for smaller models suited to local deployment and privacy. We observe that sequential decision making can stagnate when generated questions induce similar likelihoods across leading hypotheses. We address this problem by introducing belief-state-guided scoring and prompting for Bayesian interviews. The method uses the system's turn-by-turn confidence in each hypothesis to make the next-question objective explicit: discriminate among leading hypotheses while preserving the full hypothesis space and posterior, avoiding premature irreversible exclusions.

We instantiate this approach in medical diagnosis, where cost-aware sequential reasoning is critical. In this domain, a $15 symptom question may narrow the hypothesis space as effectively as a $200 lab test, but only if the system tracks which hypotheses remain plausible at that point in the interview. Further, we focus on locally deployable models (e.g., MedGemma 4B (Sellergren et al., 2025), Qwen3-VL 8B (Qwen Team, 2025)) to support privacy and show that they perform strongly on real emergency department cases.

Our paper makes the following contributions: (1) adaptive control strategies that dynamically adjust question generation and scoring based on belief dynamics, (2) cost-aware query selection that reduces expenditure without sacrificing accuracy, and (3) on real emergency department cases, 86.6% diagnostic accuracy at 31% lower cost than the best non-adaptive baseline—compared to 35.8% accuracy at nearly 4$\times$ the cost for pure LLM selection. To our knowledge, this is the first peer-reviewed empirical evaluation of LLM-mediated Bayesian experimental design, with full implementation released for reproducibility.

2 Related work

The problem of selecting informative queries has been studied extensively in statistics under the framework of Bayesian experimental design (Chaloner & Verdinelli, 1995). The Expected Information Gain (EIG) criterion, introduced by Lindley (1956), selects queries that maximize expected reduction in posterior entropy. This objective has been applied to active learning (MacKay, 1992), adaptive testing, and sequential experimental design. Recent work scales these methods using deep learning (Foster et al., 2021; Kleinegesse & Gutmann, 2020), elicits preferences and beliefs in a Bayesian manner from LLMs (Handa et al., 2024), and improves the LLM probabilistic consistency through fine-tuning (Qiu et al., 2026). SOTA prior work used EIG to guide information gathering with LLMs in 20-questions style domains (Choudhury et al., 2025).

For agents whose objective is cumulative reward, multi-armed bandits and UCB-style exploration are the natural framework (Auer et al., 2002). Our setting differs in objective: the agent gathers information about a hypothesis rather than accumulating reward over arms, and the per-question scoring depends on the evolving posterior p_t rather than on stationary per-arm rewards. UCB confidence terms are also calibrated for unbiased noise on a scalar reward, whereas LLM likelihood estimates are vector-valued and exhibit systematic biases. The empirical pattern across our experiments is consistent with this: outcomes depend on properties of the LLM's likelihood backbone rather than on exploration–exploitation calibration. We extend this line with adaptive information criteria, cost-aware queries, and

a medical-diagnosis application, focusing on failure modes that arise when LLM-generated questions induce similar likelihoods across leading hypotheses.

Bayesian networks have long been applied to medical diagnosis (Lucas et al., 2004), representing disease-symptom relationships as probabilistic graphical models. Reinforcement learning approaches have also been proposed for symptom checking (Kao et al., 2018). LLMs have shown strong performance on medical question-answering benchmarks (Singhal et al., 2023), but interactive diagnosis with LLMs raises challenges around uncertainty quantification and cost-aware decision-making that pure language modeling does not address. Further, a large agentic-AI system diagnosed 300 of the most challenging medical cases using low-cost tests (Nori et al., 2025).

Rather than using LLMs as end-to-end reasoners, recent work has explored integrating them as components in larger systems. Dohan et al. (2022) proposed viewing language models through the lens of probabilistic programming, enabling composition with explicit inference procedures. Wong et al. (2023) integrate LLMs with symbolic reasoning. Our framework extends this perspective to sequential medical decision-making, using the LLM for semantic tasks while an explicit Bayesian loop handles uncertainty and cost-aware action selection.

3 Problem formulation

We study hypothesis identification via sequential queries. An agent assumes a hypothesis space $\mathcal{H} = \{h_1, \ldots, h_N\}$. Its goal is to identify $h^* \in \mathcal{H}$ accurately and cheaply using natural-language questions and costly tests.

The agent starts with an initial context x_0 (e.g., a patient's chief complaint) and prior beliefs $p_0(h)$. The prior may be uniform or informative depending on the context. At each turn $t = 1, \ldots, T$, a query q_t is selected from a candidate set $\mathcal{Q}_t$. Queries fall into two categories: **questions**, yes/no natural-language queries posed to the respondent, each carrying a nominal cost c_{q}, and **diagnostic tests**, domain-specific actions (e.g., laboratory tests, imaging) that return structured evidence and carry procedure-specific costs $c(q_t) \gg c_{\mathrm{q}}$. After receiving the answer a_t, the system updates its belief distribution $p_t(h)$ via Bayes' rule using likelihood estimates $P(a_t \mid q_t, h)$ provided by an LLM.

We seek to select queries to identify h^* accurately while minimizing total cost: return the *maximum a posteriori* hypothesis $\hat{h}$ within T turns. That is, the policy maximizes diagnostic accuracy (i.e., $\hat{h} = h^*$) while keeping total cost low. Because each query carries a cost, the appropriate criterion for choosing among candidates is expected information per unit of cost. At each turn, the system therefore selects the query that maximizes this ratio. See Section 4 for scoring functions and adaptive mechanisms.

The LLM serves two functions in this framework. First, **query generation**: given the current belief state and conversation history, it proposes a set of candidate queries $\mathcal{Q}_t$. The prompt is conditioned on the top hypotheses and their probabilities, enabling belief-aware generation. Second, **likelihood estimation**, where for each candidate question q and hypothesis h, the LLM estimates $P(\text{Yes} \mid q, h)$, providing the conditional probabilities needed for Bayesian scoring and belief updates. Crucially, the LLM does not select queries or maintain beliefs.

4 Methods

We formulate interviewing as Bayesian sequential information gathering with LLM-generated questions and likelihood estimates. At each turn, the system maintains a posterior over hypotheses, prompts the LLM for candidate yes/no queries, estimates $P(a \mid q, h)$ for each query-hypothesis pair, and selects the next query with an explicit scoring rule. The roles are thus cleanly separated: the LLM supplies semantic candidates and likelihoods, while the Bayesian layer tracks beliefs and selects actions.

Baselines and positioning. We first define our LLM-mediated EIG baseline, which follows the core idea of Choudhury et al. (2025). As no public implementation is available, we im-

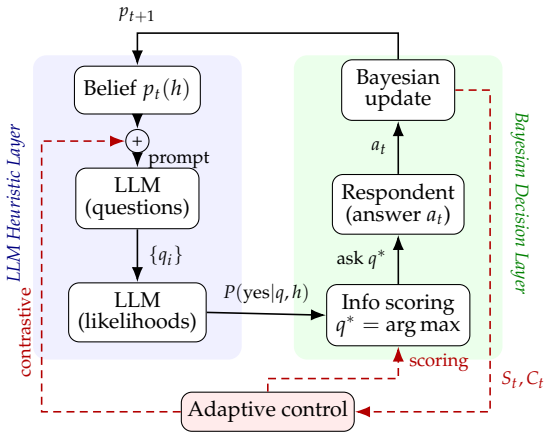


Figure 1: Proposed framework. The LLM heuristic layer (left) generates candidate questions $\{q_i\}$ and estimates likelihoods. The Bayesian decision layer (right) computes the expected information change score, selects q^*, and updates beliefs. Adaptive control (below, dashed) monitors surprise S_t and concentration C_t, injecting contrastive prompts via the $\oplus$ node and modifying scoring when stagnation is detected. While shown with LLM questions and likelihoods, we also integrate diagnostic tests with varied costs.

plement the core idea and adapt it to our sequential-diagnosis objective: the LLM generates candidate questions, estimates $P(\text{Yes} \mid q, h)$, and a Bayesian layer selects the maximum-EIG question. We also include a direct *LLM-Select* baseline, where the LLM generates candidates and chooses among them using its own judgment of informativeness (Appendix B). Our method adds belief-state-guided scoring and prompting while keeping posterior updates over the full hypothesis space. We describe the stages, adaptive rules, and control criteria below.

Query generation. At each turn t, the LLM generates k candidate yes/no queries conditioned on the current belief state. The prompt includes top hypotheses with probabilities, question history, and instructions to distinguish among candidates. Multiple candidates allow the Bayesian engine to select the most informative option.

Conditional probability estimation. The system estimates $P(\text{Yes} \mid q, h)$: the probability that the answer to q would be "yes" if h were true. Estimates are obtained by querying the LLM: "If the target hypothesis is [h], would the answer to: [q] be yes?" The probability is extracted from Yes/No token log-probabilities given by the LLM.

Question scoring. Given the LLM-generated candidate set, the Bayesian decision layer scores each candidate under the current posterior and selects the highest-scoring query. For the EIG baseline, following the expected-information criterion of Lindley (1956), the score is:

$$s_{\text{EIG}}(q) = H(p_t) - \sum_{a \in \{\text{yes,no}\}} P(a \mid q, p_t)\, H(p_{t+1}^{(q,a)}). \tag{1}$$

Here $P(a \mid q, p_t) = \sum_h p_t(h) P(a \mid q, h)$ is the marginal probability of answer a and $p_{t+1}^{(q,a)}$ is the posterior after observing answer a. In ideal settings with a fixed candidate set and calibrated likelihoods, greedy EIG selection is adaptive submodular and recovers a $(1 - 1/e)$ approximation to the offline-optimal policy (Golovin & Krause, 2011). We do not inherit this guarantee directly because the LLM produces both $\mathcal{Q}_t$ and approximate likelihoods $P(a \mid q, h)$. We therefore treat adaptive submodularity as motivation, and design our adaptive variants to address a specific failure mode that arises when the assumptions are violated. The key stretched assumption is bounded i.i.d. likelihood error: LLM-estimated likelihoods are bounded in magnitude but their errors correlate across (h, q) pairs in ways

that reflect model-specific competence patterns rather than i.i.d. perturbation, a pattern we observed in pilot studies. The empirical results in Section 5, including the model-pairing patterns, are consistent with this characterization.

When full-posterior EIG is mismatched to the decision. The terminal decision is which leading hypothesis is correct, a top-k judgment rather than a full-posterior judgment. When p_t is broad, full-posterior EIG (computed with LLM-estimated likelihoods) can spend scoring budget on low-probability hypotheses; when leading hypotheses are similar, candidate questions can induce similar likelihoods and small belief updates. Our adaptive variants therefore score on the top-k while continuing Bayesian updates over the full hypothesis space, focusing question selection on decision-relevant discriminations while preserving recovery if the leaders are wrong.

We expect the benefit to be largest when the posterior remains broad and smallest when an informative prior or pruning has already concentrated it; Section 5 reports this ordering empirically. The adaptive variants keep the same scoring interface, $s(q)$, but change the score when focused discrimination is more useful than global entropy reduction. We test two scores as belief-aware alternatives. In the following, $\mathcal{K}_t$ denotes the top-k hypotheses under p_t and let $\tilde{p}_t$ be p_t renormalized over $\mathcal{K}_t$.

Truncated EIG (s_{tEIG}) computes the information gain only over the current focused set:

$$s_{\text{tEIG}}(q) = H(\tilde{p}_t) - \sum_{a \in \{\text{yes,no}\}} P(a \mid q, \tilde{p}_t)\, H(\tilde{p}_{t+1}^{(q,a)}). \tag{2}$$

This focuses scoring on discriminating the leading hypotheses, rather than diluting the signal with low-probability candidates unlikely to affect the final decision.

Pairwise discrimination (s_{PD}) takes a more direct approach: rather than measuring entropy reduction, it explicitly measures how well a question separates hypotheses. Related to the Gini-Simpson index (Gini, 1936; Simpson, 1949) and a form of Rao's Quadratic Entropy (Rao, 2010):

$$s_{\text{PD}}(q) = \sum_{i,j \in \mathcal{K}_t, i<j} \tilde{p}_t(h_i)\, \tilde{p}_t(h_j)\, d_{ij}(q)^2, \tag{3}$$

where $d_{ij}(q) = P(\text{Yes} \mid q, h_i) - P(\text{Yes} \mid q, h_j)$ is the difference in yes-probability between h_i and h_j. It rewards queries with diverging expected answers among the top candidates. These metrics can be combined: a *blended* scorer uses $s_{\text{blend}}(q) = \alpha\, s_{\text{tEIG}}(q) + (1-\alpha)\, s_{\text{PD}}(q)$, combining truncated EIG with pairwise discrimination. As the blend is only activated when belief is focused on few candidates, using truncated EIG is natural and computationally efficient. The adaptive control mechanism (described below) can switch between metrics when stagnation is detected.

Belief update. After posing question q_t and receiving answer a_t, beliefs are updated via Bayes' rule: $p_{t+1}(h) = P(a_t \mid q_t, h) \cdot p_t(h) / \left(\sum_{h'} P(a_t \mid q_t, h') \cdot p_t(h')\right)$.

Cost-aware scoring. When queries have non-uniform costs, we incorporate cost information directly into the question-selection objective. Let $c(q)$ denote the cost of candidate question or test q. The cost-aware score is $s_{\text{cost}}(q) = s(q)/c(q)$. This ratio favors questions that yield high information per unit cost, naturally balancing diagnostic value against resource expenditure. Free-form yes/no questions are assigned a fixed nominal cost, while laboratory tests use procedure-specific estimates.

Adaptive control. The explicit belief distribution acts as a control signal for LLM behavior. We use two strategies that steer scoring or prompting toward currently relevant hypotheses while maintaining Bayesian updates over the full hypothesis space.

Strategy 1: Surprise-reactive control. This strategy detects when the interview has stalled and intervenes to break out. We measure surprise using Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) between successive belief distributions; low KL indicates that questions are no longer shifting beliefs meaningfully. To avoid intervening too eagerly,

we require the posterior to have concentrated on a small set of leading candidates before triggering. We define surprise at turn t as the KL divergence between successive beliefs:

$$S_t = D_{\mathrm{KL}}(p_t \,\|\, p_{t-1}) = \sum_{h \in \mathcal{H}} p_t(h) \log \frac{p_t(h)}{p_{t-1}(h)}, \tag{4}$$

and concentration, $C_t = \sum_{h \in \text{Top-}k(p_t)} p_t(h)$, as the posterior mass on the top-k hypotheses. Surprise-reactive control activates when $S_t < \tau_S$ for M consecutive turns and $C_t \geq \tau_C$. It deactivates when surprise exceeds a higher threshold τ_S^{exit} or concentration drops below τ_C. Then the system uses standard EIG again. When triggered, question generation is augmented with a targeted prompt that contrasts the top hypotheses. Instead of requesting generic split questions, the prompt names the leading candidates and asks for questions with opposing likely answers. For respiratory conditions, such questions can focus on markers like onset dynamics. Candidates from both prompts are scored.

Strategy 2: Truncated EIG control. This strategy continuously focuses question scoring on the current top candidates. At each turn, the truncated EIG scorer (defined in the previous section) computes information gain only over the renormalized belief distribution $\tilde{p}_t$ restricted to the top-k hypotheses. Crucially, belief updates after each answer still apply Bayes' rule over the full hypothesis space $\mathcal{H}$. This separation has an important consequence: questions are always directed at discriminating the leading candidates, but the set of leading candidates changes with evidence. If evidence shifts mass to a previously unlikely hypothesis, it enters the top-k set and subsequent questions target it.

5 Experiment

We focus the main evaluation on the clinical diagnosis setting, where query costs and belief concentration are central to the problem. For a controlled reference point, Appendix B therefore tests the same design on a non-clinical animal-identification task where exact likelihoods are available, comparing LLM-mediated EIG against a calibrated-table baseline and our adaptive variants. The comparison shows that the benefit of adaptive scoring is not domain-specific.

We evaluate our framework on a medical diagnosis task using emergency department records, where a simulated clinician must identify the correct diagnosis by asking questions and ordering tests. More specifically, we use the MIMIC-IV-ED v2.2 dataset (Johnson et al., 2023), a publicly available collection of approx. 425,000 emergency department visits from Beth Israel Deaconess Medical Center (2011–2019). Each record includes ICD diagnosis codes, triage information (vital signs, chief complaint, acuity level), and patient demographics. We preprocess and filter the data as described in Appendix C and randomly sample 1,000 visits for evaluation. All experiments ran over two weeks on an internal cluster with NVIDIA RTX 6000 Ada GPUs (48 GB VRAM).

Each diagnostic session involves two agents: a *Questioner* (the system under evaluation) that asks questions or orders medical tests to identify the diagnosis, and an *Oracle* that provides ground-truth answers. Queries are restricted to binary (yes/no) questions for tractable Bayesian updates. At each turn, the Questioner proposes three candidate questions and three candidate medical tests; the best query is selected by the scoring method. Sessions terminate when the posterior probability of the top hypothesis exceeds 0.9 or after 15 turns. Full session details are in Appendix E.

For each candidate query q, the system estimates conditional likelihoods $P(\text{Yes} \mid q, d)$ for all diseases d in the active hypothesis set (either all 100 diseases, or the focused set for adaptive/truncated methods). These likelihoods are used to compute the scoring metric (EIG, PD, or blend). After selecting the highest-scoring query, the Oracle provides the ground-truth answer: given the true diagnosis d^*, the Oracle estimates $P(\text{Yes} \mid q, d^*)$ and thresholds at 0.5 to produce a binary response. The system then performs the Bayesian update over all 100 candidate diseases using the full likelihood vector, allowing adaptive methods to recover diagnoses outside the focused set when early beliefs are incorrect. Refer to Appendix F and E for a detailed review of the information flow.

Table 1: Method comparison under an informative prior with no pruning (I-NP), $n = 1{,}000$ cases. Accuracy and C&C are percentages [95% CI]; Cost (USD) and Qs (average number of queries per session) are mean $\pm$ 95% CI half-width. Best value per column in bold.

Method	Acc. (%)	C&C (%)	Cost ($)	Qs
LLM-Select	35.8 [32.9–38.8]	11.2 [9.4–13.3]	1130 ± 77	10.3 ± 0.3
EIG	81.8 [79.3–84.1]	71.3 [68.4–74.0]	526 ± 44	7.4 ± 0.3
EIG-Cost-Aware	81.3 [78.8–83.6]	70.4 [67.5–73.1]	299 ± 31	7.6 ± 0.3
Truncated-EIG-CA	84.5 [82.1–86.6]	74.5 [71.7–77.1]	213 ± 19	7.4 ± 0.3
Truncated-PD-CA	86.1 [83.8–88.1]	76.7 [74.0–79.2]	236 ± 24	**7.2** ± 0.3
Truncated-PD-EIG-CA	**86.6** [84.3–88.6]	76.2 [73.5–78.7]	**205** ± 19	7.4 ± 0.3
Adaptive-EIG-CA	86.3 [84.0–88.3]	76.7 [74.0–79.2]	222 ± 21	7.6 ± 0.3
Adaptive-PD-CA	86.5 [84.2–88.5]	**77.0** [74.3–79.5]	248 ± 23	7.4 ± 0.3
Adaptive-PD-EIG-CA	84.6 [82.2–86.7]	74.5 [71.7–77.1]	231 ± 21	7.5 ± 0.3

We evaluate two settings of priors. Under the *informative prior*, for each candidate disease d, we prompt the LLM with the patient's triage information and demographics, and ask whether the presentation is consistent with d. We use the model-estimated probability of the "Yes" token as a raw weight w_d. Following the principle of weakly informative priors (Gelman et al., 2017), we set $P(d) = \exp(\log w_d/\tau)/\sum_{d' \in \mathcal{D}} \exp(\log w_{d'}/\tau)$ with $\tau = 2.0$, then cap the maximum prior at 0.30 and renormalize to prevent overconfidence.

Under the *uniform prior*, all 100 diseases receive equal probability. This condition simulates scenarios where initial patient information is unavailable or unreliable. We use `google/medgemma-4b-it` (Sellergren et al., 2025) throughout, for both the Questioner and the Oracle; Appendix H.5 reports a split setting that uses MedGemma-27B for Oracle answers and likelihood estimation, mitigating shared-model self-consistency concerns.

We compare nine methods spanning three design dimensions: scoring metric, cost-awareness, and adaptive control (Table 2, Appendix D). *EIG* selects a question or test by expected information gain over the full hypothesis space, whereas *LLM-Select* asks the LLM to score each candidate by the probability that it is the best next query. *Cost-aware variants* divide the score by query cost, assigning a fixed cost of $15 per question and using LLM-estimated typical U.S. out-of-pocket prices for medical tests. We evaluate two adaptive approaches. *Truncated* variants restrict question generation and scoring to the focused set (top hypotheses covering $\geq$95% of posterior mass), except when the prior is uniform. *Adaptive* variants dynamically switch between standard and focused modes (entry: $S_t < 0.15$ and top 20% hold $\geq$90% mass; exit: $S_t > 0.40$ or loss of concentration). For completeness, we also evaluate *pruning*, i.e., removing the lowest-posterior diagnoses after each turn, as an orthogonal intervention. All adaptive and truncated variants use cost-aware scoring.

We report *Accuracy* (top diagnosis matches ground truth), *Correct & Converged* (C&C; correct and reaches 90% confidence before maximum turns), *Average Cost* (total question and test costs per session), and *Average Questions* (turns per session). All metrics include 95% confidence intervals. Additional metrics including session duration and final confidence are reported in Appendix G. We evaluate our method under four conditions: informative and uniform priors, each with and without pruning (Appendix D). We focus here on the Informative Prior and No Pruning (I-NP), the most realistic and conservative setting.

Score-based methods dramatically outperform LLM-Select. LLM-Select achieves only 35.8% [32.9–38.8%] accuracy, compared to 81.3% or higher for every score-based method (Table 1). The cost gap is equally pronounced: LLM-Select averages $1,130 per session, roughly double the cost of EIG ($526) and five times the cost of the best-performing adaptive methods. Its C&C rate of 11.2% is far below the 70%+ achieved by score-based methods, indicating that it rarely reaches a confident, correct diagnosis within the allotted turns.

Cost-aware scoring substantially reduces expenditure without sacrificing accuracy. Comparing EIG to EIG-Cost-Aware, cost drops from $526 to $299 (43% reduction) with negligible accuracy change (81.8% vs. 81.3%). Adaptive-family cost-aware methods reduce costs

further while improving accuracy. Truncated-PD-EIG-Cost-Aware achieves the highest accuracy at 86.6% [84.3–88.6%] with only $205 average cost, a 61% reduction relative to plain EIG. Across all adaptive and truncated methods, cost falls in the range $205–$248 versus $299–$526 for the non-adaptive baselines.

Methods incorporating pairwise discrimination (PD), alone or blended with EIG, achieve the highest accuracy. The top performers are Truncated-PD-EIG-CA (86.6%), Adaptive-PD-CA (86.5%), Adaptive-EIG-CA (86.3%), and Truncated-PD-CA (86.1%). Adaptive-EIG-CA reaches comparable accuracy without PD, indicating that adaptive control itself contributes substantially to the improvement over plain EIG. EIG-only methods cluster at 81.3–81.8%.

Comparing the best adaptive-family method against EIG-Cost-Aware across the four conditions (Appendix G), the accuracy gain is largest under the broadest posterior and shrinks to zero under the narrowest: U-NP +7.4 percentage points (86.6% vs 79.2%), I-NP +5.3pp (86.6% vs 81.3%), U-P +1.0pp (89.0% vs 88.0%), and I-P $\approx$ 0pp (87.9% vs 88.0%). Cost reductions follow a similar pattern, with U-NP reducing average session cost from $634 for plain EIG to $220–$291 for adaptive-family variants. This U-NP $>$ I-NP $>$ U-P $>$ I-P pattern is consistent with the prediction of Section 4: adaptive-family scoring helps most when neither an informative prior nor pruning has already concentrated the posterior, and offers little once the candidate set is concentrated by other means. Additional robustness studies (Appendix H) support the role of explicit scoring while showing that adaptive gains are model-dependent.

6 Discussion and limitations

We show our hybrid LLM-Bayesian system with adaptive cost-aware information criteria supports cost-aware medical diagnosis beyond pure LLM systems and the prior state of the art. Cost-awareness serves accurate diagnosis under resource constraints: it is a means of allocating limited testing well, not of reducing care.

Unlike pure LLM systems, the framework exposes its reasoning. At each turn, the clinician can inspect the current probability distribution over diagnoses, the EIG scores of candidate questions, and the rationale for test recommendations. This interpretability is important for clinical adoption, where opaque AI recommendations face justified skepticism.

We intentionally use small LLMs to show that the Bayesian layer can support robust belief updates without large models, but the tradeoff is limited world knowledge. In a medical side experiment with a domain-general Qwen3 30B-class model ($n = 200$; Appendix H, Table 12), the best-performing method on MedGemma-4B (Truncated-PD-EIG-Cost-Aware) drops below EIG-Cost-Aware (76.0% vs 80.5%, overlapping CIs) and calibration was markedly worse than MedGemma-4B (overconfidence gap +15.7% to +19.2% versus +5.1% to +8.1%; Cohen's $d \approx 0.85$ versus 1.25). On this task, calibration seems to be driven by domain specialization more than parameter count. A complementary split Questioner–Oracle/likelihood study (Appendix H.5) substitutes the larger MedGemma-27B for Oracle answers and likelihood estimation while retaining the MedGemma-4B Questioner; here the best truncated cost-aware variant reaches 87.0% vs 82.0% for EIG-Cost-Aware, consistent with adaptive's lift tracking the calibration of the likelihood backbone rather than whether the Questioner and Oracle share a model. This matches the framing in Section 4: adaptive scoring focuses on leading hypotheses, helping when the LLM's posterior over leaders is approximately right and offering little advantage otherwise. Thus, adaptive scoring depends on the LLM–task pairing rather than model size alone.

The framework assumes a discrete, enumerable hypothesis space, so rare conditions outside the initial set cannot be diagnosed. Future work should explore dynamic hypothesis expansion, in which the candidate set is revised mid-session when early answers fit none of the existing hypotheses well. Second, the LLM's conditional probability estimates may be miscalibrated; integrating curated clinical knowledge bases could improve reliability.

Adaptive control can occasionally hurt performance. A mitigation is to *expand* rather than *replace*: add discriminative candidates while retaining EIG-optimal questions. We omit this

from the main framework because the simpler approach improves performance on average, but if per-session guarantees are needed, this fallback can be added directly (Appendix A).

The current implementation restricts queries to yes/no questions, where answers map directly to likelihoods $P(\text{Yes}|q, h)$. Extending to open-ended questions would require the LLM to translate free-form responses into likelihoods $P(\text{answer}|h)$. Calibrating this interpretation layer for richer evidence types remains an important direction for future work.

References

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Deepro Choudhury, Sinead Williamson, Adam Goliński, Ning Miao, Freddie Bickford Smith, Michael Kirchhof, Yizhe Zhang, and Tom Rainforth. Bed-llm: Intelligent information gathering with llms and bayesian experimental design. *arXiv*, 2025. URL https://arxiv.org/abs/2508.21184.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A. Saurous, Jascha Sohl-Dickstein, Kevin Murphy, and Charles Sutton. Language model cascades. *arXiv preprint arXiv:2207.10342*, 2022.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *International Conference on Machine Learning*, pp. 3384–3395. PMLR, 2021.

Andrew Gelman, Daniel Simpson, and Michael Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555, 2017.

Georgi Gerganov. llama.cpp. https://github.com/ggml-org/llama.cpp, 2023.

Corrado Gini. On the measure of concentration with special reference to income and statistics, colorado college publication. *General series*, 208:73–79, 1936.

Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42: 427–486, 2011. doi: 10.1613/jair.3278.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.02207.

Kunal Handa, Yarin Gal, Ellie Pavlick, Noah Goodman, Jacob Andreas, Alex Tamkin, and Belinda Z Li. Bayesian preference elicitation with language models. *arXiv preprint arXiv:2403.05534*, 2024.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, 2022. URL https://arxiv.org/abs/2207.05608.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. MIMIC-IV-ED. *PhysioNet*, January 2023. doi: 10.13026/5ntk-km72. URL https://doi.org/10.13026/5ntk-km72. Version 2.2.

Hao-Cheng Kao, Kai-Fu Tang, and Edward Y. Chang. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Steven Kleinegesse and Michael U. Gutmann. Bayesian experimental design for implicit models by mutual information neural estimation. In *International Conference on Machine Learning*, pp. 5316–5326. PMLR, 2020.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Alexander K Lew, Tan Zhi-Xuan, Gabriel Grand, and Vikash K Mansinghka. Sequential monte carlo steering of large language models using probabilistic programs. *arXiv preprint arXiv:2306.03081*, 2023.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.13382.

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

Peter J. F. Lucas, Linda C. van der Gaag, and Ameen Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, 30(3):201–214, 2004.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

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.

Linlu Qiu, Fei Sha, Kelsey Allen, Yoon Kim, Tal Linzen, and Sjoerd van Steenkiste. Bayesian teaching enables probabilistic reasoning in large language models. *Nature Communications*, 2026.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

C. R. Rao. Quadratic entropy and analysis of diversity. *Sankhya: The Indian Journal of Statistics*, 72A:70–80, 2010.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

E. H. Simpson. Measurement of diversity. *Nature*, 163, 1949.

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. doi: 10.1038/s41586-023-06291-2.

Lionel Wong, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. From word models to world models: Translating from natural language to the probabilistic language of thought. In *CoRR*, volume abs/2306.12672, 2023. URL https://doi.org/10.48550/arXiv.2306.12672.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning–a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Grosse. Probabilistic inference in language models via twisted sequential monte carlo. *arXiv preprint arXiv:2404.17546*, 2024.

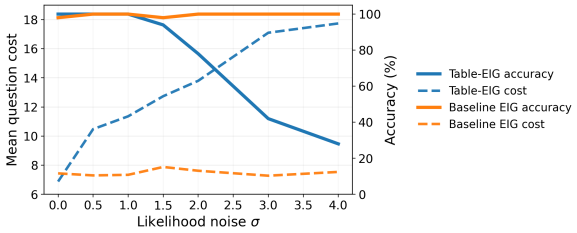


Figure 2: AwA2 likelihood-noise sweep for Table-EIG (blue) and baseline EIG (orange). Solid lines show accuracy; dashed lines show mean question cost, counting failures as 20 questions.

Figure 3: AwA2 likelihood-noise sweep for Table-EIG (blue) and LLM-mediated EIG (orange), over 50 sessions (one per animal class). The horizontal axis is the standard deviation of Gaussian noise added to likelihoods in logit space. Solid lines (left axis) show accuracy; dashed lines (right axis) show the mean number of questions per session, with sessions that fail to reach 90% confidence counted as 20.

A Data, codes, and models

Our reproducible experimental code is available at `https://github.com/hongtaoh/bayesian_eig_colm`.

We use MIMIC-IV-ED v2.2 (Johnson et al., 2023) under the PhysioNet Credentialed Health Data License 1.5.0 and the associated Data Use Agreement; access requires credentialing, required training, and agreement to the PhysioNet terms.

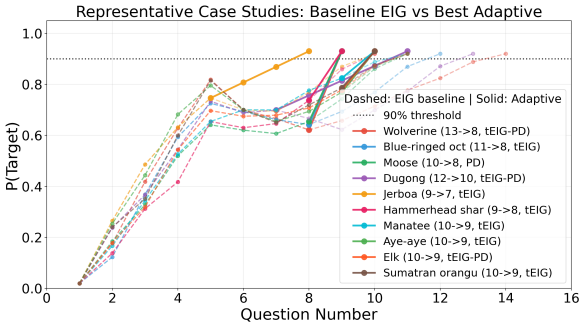
We use Animals with Attributes 2 (AwA2) (Xian et al., 2018) from the official dataset release. The image archive includes individual license files and whose webpage states that images were included only when licensed for free use and redistribution. We do not redistribute the AwA2 predicate matrix; users are instructed to download it from the official AwA2 page.

MedGemma models (Sellergren et al., 2025) are governed by the Health AI Developer Foundations terms of use. The Qwen models used in our experiments and robustness studies, including Qwen3-VL 8B (Qwen Team, 2025), Qwen3-4B, and Qwen3-30B, are distributed under the Apache 2.0 license. We use `llama.cpp` (Gerganov, 2023) for local inference under the MIT license.

B Experiment: General-domain validation

We evaluate Table-EIG and LLM-mediated EIG on the 50-class AwA2 dataset (Xian et al., 2018) with a uniform prior, running one session per animal. Each session ends when the top hypothesis reaches 90% confidence or after 20 questions. Table-EIG selects from the fixed attribute bank, while LLM-mediated EIG uses Qwen3-VL 8B to generate yes/no questions and estimate $P(\text{Yes} \mid q, h)$ for candidate questions and animals. We test both methods in the clean setting and with controlled likelihood noise applied in logit space (see Appendix A). This comparison motivates our adaptive-control setting by testing LLM-mediated EIG beyond fixed question banks, with questions and likelihoods generated on demand.

We ran all general domain validation experiments using local inference with `llama.cpp` (Gerganov, 2023). A single NVIDIA RTX 3090 or RTX 5090 GPU was sufficient to host the 8B model used for question generation and likelihood estimation, with remaining VRAM used for parallel runs. We use the same Qwen3-VL 8B interview setup to test whether adaptive scoring improves the LLM-only regime beyond baseline EIG. For the switch-based adaptive runs, the adaptive switch triggers when surprise falls below $\tau_S = 0.15$ and the top 10% of candidates hold at least $\tau_C = 0.95$ of the probability mass. $\tau_S^{\text{exit}} = 0.4$ is the surprise threshold for exiting focused mode.



	EIG	tEIG	Blend	PD
Avg. Q	7.75	7.63	7.56	7.60
Succ.	100%	100%	100%	100%
Div.	–	30	30	30
W/L	–	13/5	13/2	14/2
Saved	–	+12	+19	+15
p	–	.07	.004	.03

Figure 4: Left: confidence trajectories, with dashed baseline EIG and solid adaptive paths after divergence. Right: uniform-prior animal summary. W/L counts adaptive wins/losses against baseline s_{EIG} on the same target; p-values use a Wilcoxon signed-rank test.

Replay comparison methodology. To ensure fair comparison, we use trajectory replay. We first run baseline EIG and record the full trajectory: questions, answers, belief states, and cached likelihoods. Adaptive strategies traverse this trajectory turn by turn. While the adaptive scorer remains in EIG mode, replay proceeds deterministically using the baseline's cached questions and belief updates. When the scorer triggers a mode switch, cached candidates plus newly generated discrimination questions are re-scored. If the top-scoring question differs from the baseline's choice, the trajectory diverges and subsequent turns use fresh generation. A shared likelihood cache keeps $P(\text{Yes} \mid q, h)$ identical across strategies until divergence, eliminating LLM stochasticity before that point.

Results For AwA2 setting, LLM-mediated EIG nearly matches the calibrated Table-EIG reference, reaching 98% success with 7.44 mean questions versus Table-EIG's 100% success with 6.88 questions. Under likelihood noise, Table-EIG degrades: success falls to 78% with 13.80 questions at $\sigma = 2$ and to 28% with 17.74 questions at $\sigma = 4$. LLM-mediated EIG remains stable, maintaining 100% success with 7.54 questions at $\sigma = 4$. These results show that LLM-generated questions and likelihoods can support effective Bayesian interviewing without relying on a fixed set of predefined questions, and that the qualitative gap between Table-EIG (whose perturbed likelihoods satisfy the bounded i.i.d. error condition by construction) and LLM-EIG (whose errors do not, Section 4) is itself diagnostic of where the standard adaptive-submodularity guarantee applies and where it does not.

We next evaluate the adaptive component of Experiment 1 in the LLM-only setting. After posterior concentration, adaptive control shifts scoring toward leading-candidate discrimination. In the replay experiment, 30 of 100 sessions triggered adaptive control.

Figure 4 shows that adaptive variants preserve 100% success and reduce average question count modestly: the blended $s_{\text{tEIG}-\text{PD}}$ scorer performs best overall (13W/2L, +19 turns saved), while s_{tEIG} wins 13 sessions but loses 5 (+12 turns saved), with wins typically saving more turns than losses add. Losses reflect the trade-off between focused questioning and EIG's broader objective: a question may separate current top candidates yet be uninformative for the true target. We omit fallback checks here to evaluate the lightweight adaptive mechanism directly; in the more complex clinical setting of Experiment 2, the same no-fallback design yields a clear gain. We discuss mitigation strategies in Section 6.

In this domain, EIG is already near-optimal and posterior concentration is rapid; adaptive control yields modest but statistically significant gains, consistent with the prediction in Section 4 that adaptive's benefit grows with posterior breadth. Experiment 2 tests this prediction directly on a real clinical dataset across four conditions that vary how broad the posterior remains throughout a session.

LLM select vs. LLM-EIG baseline. The preliminary comparison supports the baseline choice discussed in Section 4: LLM-mediated EIG is the main baseline, while direct LLM

selection is retained as a simpler secondary ablation. The code and target list for this comparison are included in the released experimental code.

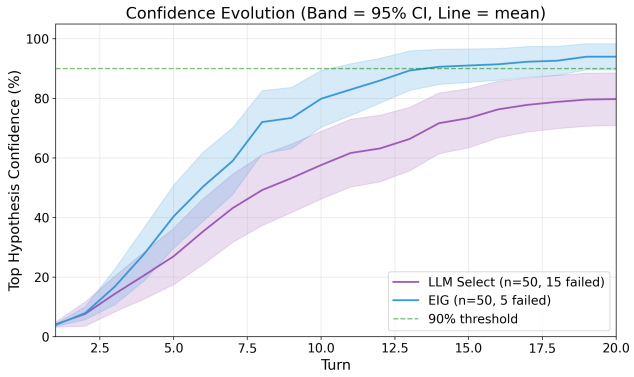


Figure 5: Mean top-hypothesis confidence (95% CI bands) over 50 disease sessions. LLM selection averages 10.7 questions vs EIG's 8.7, but LLM selection has higher variance ($\sigma = 4.8$ vs. $\sigma = 4.2$) and 3 failures at the imposed 20-question ceiling. EIG achieves 90% success vs LLM selection's 70%.

C MIMIC data processing

Diagnosis selection. Each ED stay may have multiple diagnoses recorded in the `diagnosis` table, ordered by clinical relevance. We retained only the primary diagnosis (`seq_num = 1`) for each stay, yielding 423,989 visits from 205,129 unique patients across 9,491 unique ICD titles.

Filtering non-disease codes. We excluded ICD codes that do not represent diseases: ICD-10 codes beginning with R (symptoms), Z (status), or V–Y (external causes), and ICD-9 codes beginning with E (external causes), V (status), or 78–79 (general symptoms). After standardizing titles to title case, 8,635 unique disease names remained, covering 275,114 visits.

Disease consolidation. We selected the 500 most frequent diseases, which account for 74.8% of all visits. Because many ICD titles refer to the same condition (e.g., "Pneumonia, Organism Unspecified" and "Pneumonia, Unspecified Organism"), we used Claude to map these 500 titles to 238 canonical disease labels. We then retained only the 100 most frequent canonical labels, covering 176,852 visits (85.6% of visits with top-500 diagnoses).

Data completeness filtering. We merged the filtered diagnoses with the `triage` and `edstays` tables to obtain vital signs and demographics. We excluded visits missing any of the following: chief complaint, temperature, heart rate, respiratory rate, oxygen saturation, systolic and diastolic blood pressure, pain score, or acuity level. This yielded 157,152 complete visits.

Final sample. We randomly sampled 1,000 visits (`seed = 77`), representing 99 unique disease labels, for use in our experiments.

D Experimental details

Method	Scoring	Cost-Aware	Control
LLM-Select	LLM	–	–
EIG	EIG	No	Standard
EIG-Cost-Aware	EIG	Yes	Standard
Truncated-EIG-CA	EIG	Yes	Truncated
Truncated-PD-CA	PD	Yes	Truncated
Truncated-PD-EIG-CA	Blend	Yes	Truncated
Adaptive-EIG-CA	EIG	Yes	Adaptive
Adaptive-PD-CA	PD	Yes	Adaptive
Adaptive-PD-EIG-CA	Blend	Yes	Adaptive

Table 2: Methods evaluated in the medical diagnosis experiment. CA = Cost-Aware. Blend = $0.5 \cdot \text{EIG} + 0.5 \cdot \text{PD}$.

	No Pruning	Pruning
Uniform Prior	U-NP	U-P
Informative Prior	**I-NP**	I-P

Table 3: Four experimental conditions. I-NP (bold) represents the most typical clinical scenario; other conditions address settings with limited patient information or computational constraints.

E Diagnostic session workflow

This appendix describes the diagnostic session workflow in detail, including the exact prompts used at each step. Figure 6 provides a visual overview, followed by step-by-step descriptions.

This section describes each step of a diagnostic session, specifying inputs and prompts. Variable placeholders are shown in `{braces}`.

E.1 Step 0: Compute initial prior

Inputs: Patient information (chief complaint, vital signs, demographics, acuity level); list of candidate diseases.

For the uniform prior condition, all diseases receive equal probability $P(d) = 1/|\mathcal{D}|$. For the informative prior condition, we query the LLM for each disease $d \in \mathcal{D}$:

```
Patient presentation: {patient_info}
Question: Is this presentation consistent with {disease}? Answer "Yes" or "No":
```

We extract the probability of the "Yes" token as a raw weight w_d, apply temperature-scaled softmax ($\tau = 2.0$), and cap the maximum prior at 0.30.

E.2 Step 1: generate candidate questions

Inputs: Current belief distribution, Q&A history.

Patient information is not provided at this step. Standard EIG-based methods show the full current belief. Truncated methods, except on the initial uniform prior, and adaptive methods while in focused mode show only the top diseases covering $\geq$95% of posterior mass.

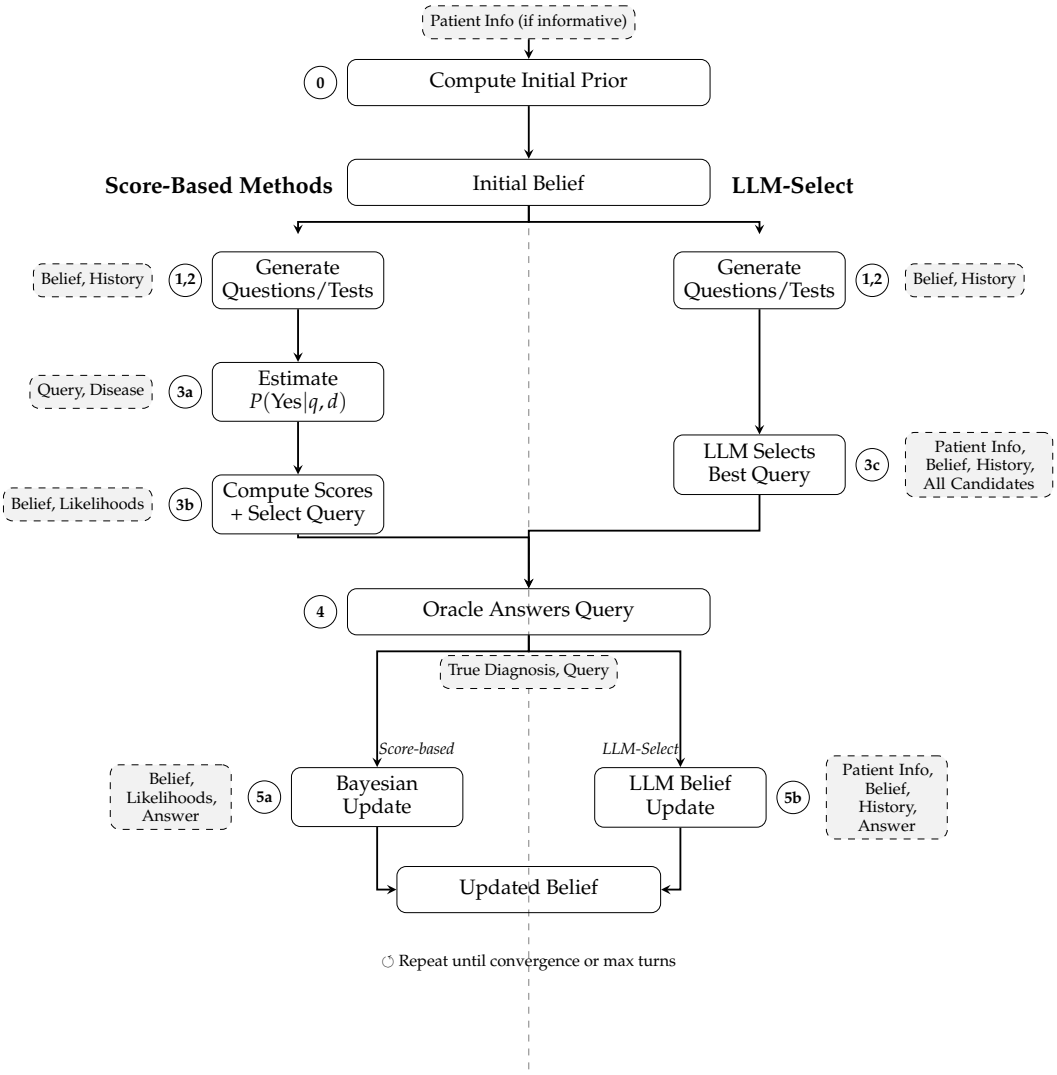


Figure 6: Workflow comparison between score-based methods (left) and LLM-Select (right). Circled numbers correspond to steps described in Section E. Dashed boxes indicate inputs to each step. Score-based methods estimate likelihoods $P(\text{Yes}|q,d)$ using only query and disease names (Step 3a), then compute explicit scores (Step 3b). LLM-Select instead evaluates candidates directly using patient context (Step 3c). Score-based methods update beliefs via Bayes' rule (Step 5a), while LLM-Select re-estimates beliefs through LLM queries (Step 5b).

```
Generate {k} yes/no questions to identify a medical diagnosis.
Current belief (diagnosis : probability): {diseases_with_probs}
Previous questions and answers, or tests and results: {history}
Requirements: - Each question must have a clear yes/no answer - Do not repeat previous
questions - The patient should be able to answer the question without any medical
tests; or that the question is about the result of a previous test shown above.
Output as JSON list: ["question 1", "question 2", ...]
```

E.3 Step 2: generate candidate tests

Inputs: Current belief distribution, Q&A history.

```
Propose {k} medical tests to identify a diagnosis.
Current belief (diagnosis : probability): {diseases_with_probs}
Previous questions with answers, and previous tests with results: {history}
Requirements: - Do not repeat previous tests - Wrap test names with <test> </test>,
e.g., <test>MRI</test>
Output as JSON list: ["<test>TEST1</test>", "<test>TEST2</test>", ...]
```

E.4 Step 3a: estimate likelihoods (score-based methods)

For score-based methods, we estimate $P(\text{Yes}|q, d)$ for each candidate query and each disease.

For questions—Inputs: Question text, disease name.

```
Given the true diagnosis is {disease}, answer this question:
Question: {question}
Answer with ONLY 'YES' or 'NO':
```

For tests—Inputs: Test name, disease name.

```
Disease: {disease} Test: {test_name}
Question: Is this test result typically POSITIVE for this disease? Answer with ONLY
'YES' or 'NO':
```

We extract the probability of the "Yes" token as the likelihood estimate. These prompts use only the query and disease—no patient information, history, or belief distribution.

E.5 Step 3b: compute scores and select query (score-based methods)

Inputs: Current belief distribution, likelihood matrix $P(\text{Yes}|q, d)$ for all queries and diseases in the active scoring set.

Using the belief distribution and likelihoods, we compute EIG, PD, or blended scores as described in Section 4. For cost-aware variants, we also estimate test costs:

```
What is a reasonable estimated out-of-pocket cost, in the United States, for a patient
WITHOUT insurance, for the medical test "{test_name}"?
If prices vary, give a single representative average estimate.
Output ONLY one integer. No words. No punctuation. No explanation. No dollar sign.
```

The query with the highest score (or highest score-to-cost ratio for cost-aware methods) is selected.

E.6 Step 3c: LLM query selection (LLM-Select only)

Inputs: Patient information, current belief distribution, Q&A history, list of all candidate queries.

Unlike score-based methods, LLM-Select provides the full patient context and asks the LLM to evaluate each candidate:

```
Patient presentation: {patient_info}
Previous questions with answers, and tests with results: {history}
Current belief (diagnosis : probability): {all_diseases_with_probs}
All candidate questions/tests: {numbered_list_of_candidates}
Evaluating candidate #{idx}: {candidate}
Given all the candidates above, is THIS candidate (#{idx}: {candidate}) the BEST
choice for making a diagnosis quickly? Answer with ONLY 'YES' or 'NO':
```

The candidate with the highest "Yes" probability is selected.

E.7 Step 4: oracle answers query

Inputs: True diagnosis, selected query.

The Oracle simulates a patient with the true diagnosis. For ordinary questions, it uses the same prompt as likelihood estimation (Step 3a), but with the true diagnosis:

```
Given the true diagnosis is {true_diagnosis}, answer this question:
Question: {selected_query}
Answer with ONLY 'YES' or 'NO':
```

For selected tests in score-based methods, the answer is obtained from the test-positive likelihood prompt:

```
Disease: {true_diagnosis} Test: {test_name}
Question: Is this test result typically POSITIVE for this disease? Answer with ONLY
'YES' or 'NO':
```

The response is converted to a probability using token log-likelihoods and thresholded at 0.5 to produce a binary answer (Yes/No for questions, Positive/Negative for tests).

E.8 Step 5a: bayesian belief update (score-based methods)

Inputs: Current belief distribution, likelihood vector $P(\text{Yes}|q, d)$ for the selected query, Oracle's binary answer.

The posterior is computed via Bayes' rule:

$$P(d|\text{answer}) = \frac{P(\text{answer}|q, d) \cdot P(d)}{\sum_{d'} P(\text{answer}|q, d') \cdot P(d')} \tag{5}$$

where $P(\text{Yes}|q, d)$ comes from likelihood estimation, and $P(\text{No}|q, d) = 1 - P(\text{Yes}|q, d)$.

In non-pruning runs, even adaptive/truncated methods that focus on top diseases during scoring perform the Bayesian update over all 100 diseases using the full likelihood vector. This preserves the ability to recover if early beliefs were incorrect. In pruning runs, the same update is performed over the retained disease set after previously pruned diagnoses have been removed.

E.9 Step 5b: LLM belief update (LLM-Select only)

Inputs: Patient information, current belief distribution, Q&A history (including the latest answer).

LLM-Select does not use Bayes' rule. Instead, it re-estimates the belief distribution by querying the LLM for each disease:

```
Patient presentation: {patient_info}
Current belief (diagnosis : probability): {all_diseases_with_probs}
Previous questions with answers, and previous tests with results: {history}
Based on all the evidence above, does this patient have {disease}? Answer with ONLY
'YES' or 'NO':
```

The "Yes" probabilities are normalized to form the updated belief distribution.

E.10 Termination

The session terminates when either (1) the posterior probability of the top-ranked disease exceeds 0.9, or (2) the maximum of 15 turns is reached. The top-ranked disease at termination is the system's diagnosis.

F Summary of information flow

Table 4 summarizes what information is available at each step.

Table 4: Information available at each step of the diagnostic process. ✓ indicates the information is used; – indicates it is not used. Patient Info includes chief complaint, vital signs, demographics, and acuity. Q&A History includes all previous questions/tests and their binary answers. Belief refers to the current probability distribution over diseases. True Diagnosis is the ground-truth label, available only to the Oracle.

Step	Patient Info	Q&A History	Belief	True Dx
Initialization				
Compute informative prior	✓	–	–	–
Each turn (all methods except LLM-Select)				
Generate candidate questions	–	✓	✓	–
Generate candidate tests	–	✓	✓	–
Estimate $P(\text{Yes}\|q, d)$ for questions	–	–	–	–
Estimate $P(\text{Pos}\|t, d)$ for tests	–	–	–	–
Compute scores (EIG/PD)	–	–	✓	–
Select best query	–	–	–	–
Oracle answers query	–	–	–	✓
Bayesian belief update	–	–	✓	–
Each turn (LLM-Select only)				
Generate candidate questions	–	✓	✓	–
Generate candidate tests	–	✓	✓	–
Select best query	✓	✓	✓	–
Oracle answers query	–	–	–	✓
LLM belief update	✓	✓	✓	–

G Detailed results

G.1 Effect of pruning

Pruning irreversibly removes the bottom 5% of retained candidates by posterior rank at the end of each turn. This concentrates later scoring and updating, but carries inherent risk because recovery becomes impossible if the true diagnosis is pruned.

Pruning benefits EIG and EIG-Cost-Aware substantially. With informative priors, EIG-Cost-Aware improves from 81.3% accuracy and $299 cost (I-NP) to 88.0% accuracy and $192 cost (I-P), becoming the top-performing method in that condition. With uniform priors, Adaptive-PD-EIG-Cost-Aware achieves the highest accuracy at 89.0% in U-P. Pruning narrows the gap between EIG-based and adaptive methods, since pruning itself provides a form of candidate focusing. Adaptive methods combined with pruning remain competitive but their advantage over EIG-Cost-Aware is smaller than in the no-pruning setting.

In the detailed result tables, CA abbreviates Cost-Aware.

Given the high-stakes nature of medical diagnosis, we recommend the non-pruning adaptive approaches as the default choice. They achieve strong accuracy (up to 86.6%) with substantial cost reduction while preserving the ability to recover from early errors. Pruning may be considered in time-critical scenarios where faster convergence is prioritized.

G.2 Not prune with informative priors

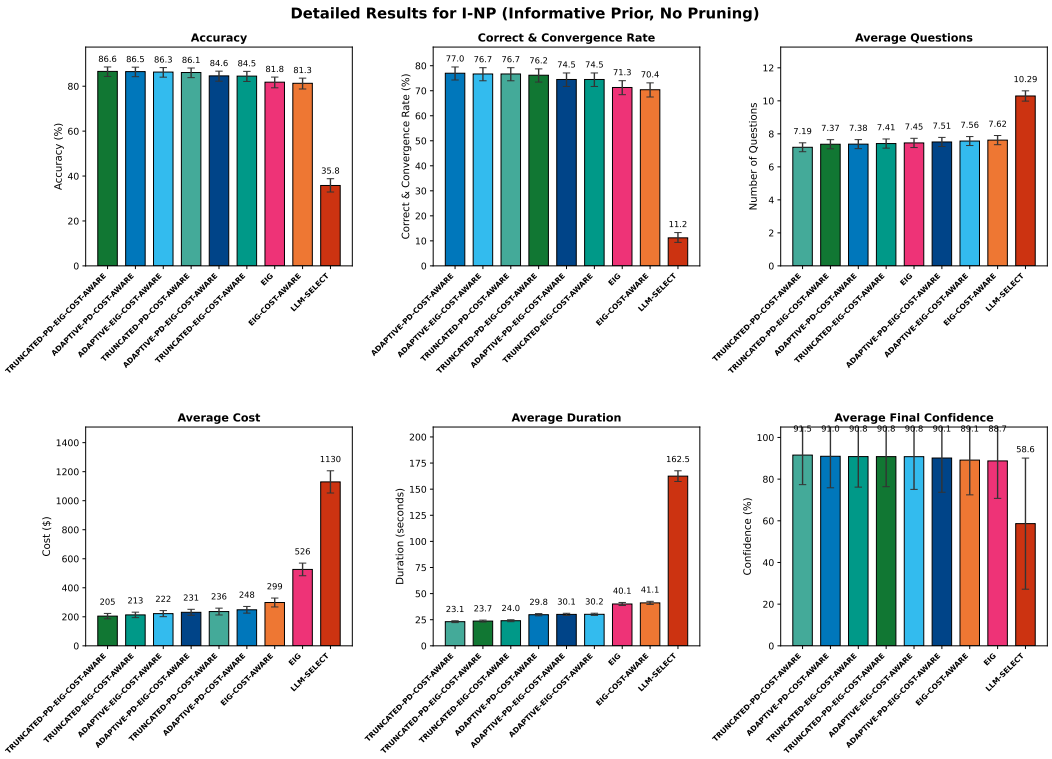


Figure 7: **Performance metrics for Informative Prior with No Pruning (I-NP).** This condition represents the standard clinical workflow where patient history informs the start. Adaptive and truncated strategies maintain high accuracy (∼86.6%) while reducing computational time and diagnostic cost relative to EIG baselines. LLM-Select fails to achieve reliable accuracy.

Table 5: Summary results for I-NP (Informative Prior, No Pruning, $n = 1{,}000$). Values show mean with 95% CI in brackets.

Method	Acc (%)	C&C (%)	Questions	Cost ($)	Duration (s)
Truncated-PD-EIG-CA	86.6 [84.3–88.6]	76.2 [73.5–78.7]	7.4 ± 0.3	205 ± 19	23.7 ± 0.9
Adaptive-PD-CA	86.5 [84.2–88.5]	77.0 [74.3–79.5]	7.4 ± 0.3	248 ± 23	29.8 ± 1.1
Adaptive-EIG-CA	86.3 [84.0–88.3]	76.7 [74.0–79.2]	7.6 ± 0.3	222 ± 21	30.2 ± 1.0
Truncated-PD-CA	86.1 [83.8–88.1]	76.7 [74.0–79.2]	7.2 ± 0.3	236 ± 24	23.1 ± 0.9
Adaptive-PD-EIG-CA	84.6 [82.2–86.7]	74.5 [71.7–77.1]	7.5 ± 0.3	231 ± 21	30.1 ± 1.1
Truncated-EIG-CA	84.5 [82.1–86.6]	74.5 [71.7–77.1]	7.4 ± 0.3	213 ± 19	24.0 ± 0.9
EIG	81.8 [79.3–84.1]	71.3 [68.4–74.0]	7.4 ± 0.3	526 ± 44	40.1 ± 1.5
EIG-CA	81.3 [78.8–83.6]	70.4 [67.5–73.1]	7.6 ± 0.3	299 ± 31	41.1 ± 1.5
LLM-Select	35.8 [32.9–38.8]	11.2 [9.4–13.3]	10.3 ± 0.3	1130 ± 77	162.5 ± 5.1

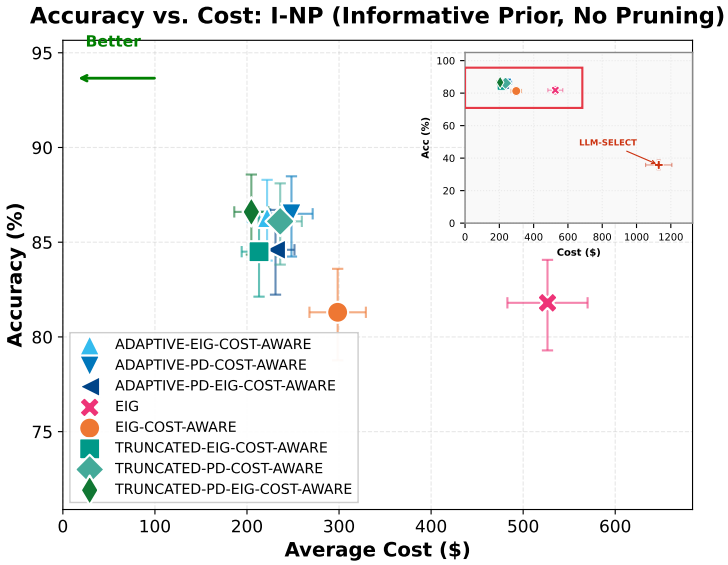


Figure 8: Experiment results: cost-performance tradeoff for different methods. Adaptive and truncated methods are more accurate and cost-efficient than EIG, EIG-Cost-Aware, and LLM-Select baselines in this condition. Error bars are 95% confidence intervals.

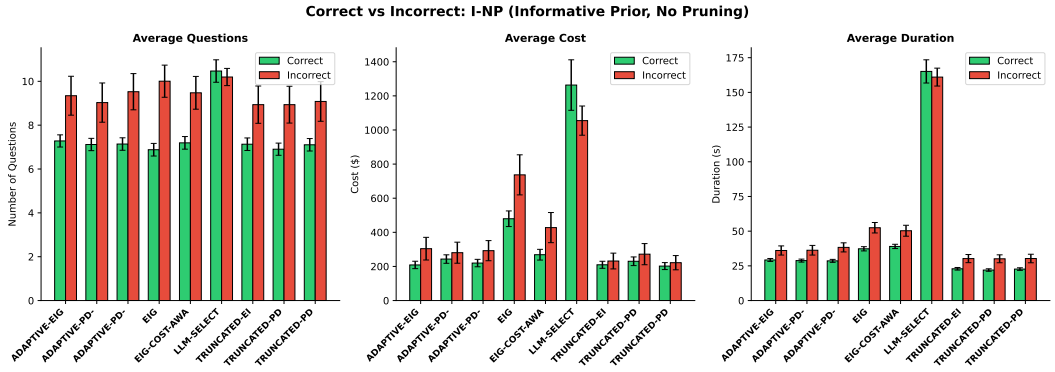


Figure 9: **Resource consumption for Correct vs. Incorrect diagnoses (I-NP).** A comparison of the questions asked and costs incurred when the system succeeds versus when it fails. This highlights whether errors stem from premature termination (low cost/questions) or stagnation (max cost/questions).

Table 6: Calibration analysis for I-NP (Informative Prior, No Pruning, $n = 1{,}000$). Analysis restricted to high-confidence predictions (confidence ≥ 0.9). "% High Conf" shows the proportion of predictions in this region. "Avg Conf" and "Accuracy" are computed within this high-confidence subset. Gap = Avg Conf − Accuracy; positive values indicate overconfidence.

Method	% High Conf	Avg Conf	Accuracy	Gap	Status
Truncated-PD-EIG-CA	85.9%	95.6%	90.6%	+5.1%	Overconfident
Adaptive-EIG-CA	86.7%	96.0%	90.7%	+5.3%	Overconfident
Adaptive-PD-CA	86.5%	95.9%	90.5%	+5.4%	Overconfident
Truncated-PD-CA	87.8%	96.1%	90.1%	+6.0%	Overconfident
Truncated-EIG-CA	86.5%	95.6%	89.1%	+6.4%	Overconfident
Adaptive-PD-EIG-CA	85.0%	96.0%	89.4%	+6.5%	Overconfident
EIG	80.9%	96.4%	89.2%	+7.1%	Overconfident
EIG-CA	81.4%	95.8%	87.7%	+8.1%	Overconfident
LLM-Select	36.5%	97.6%	31.0%	+66.6%	Overconfident

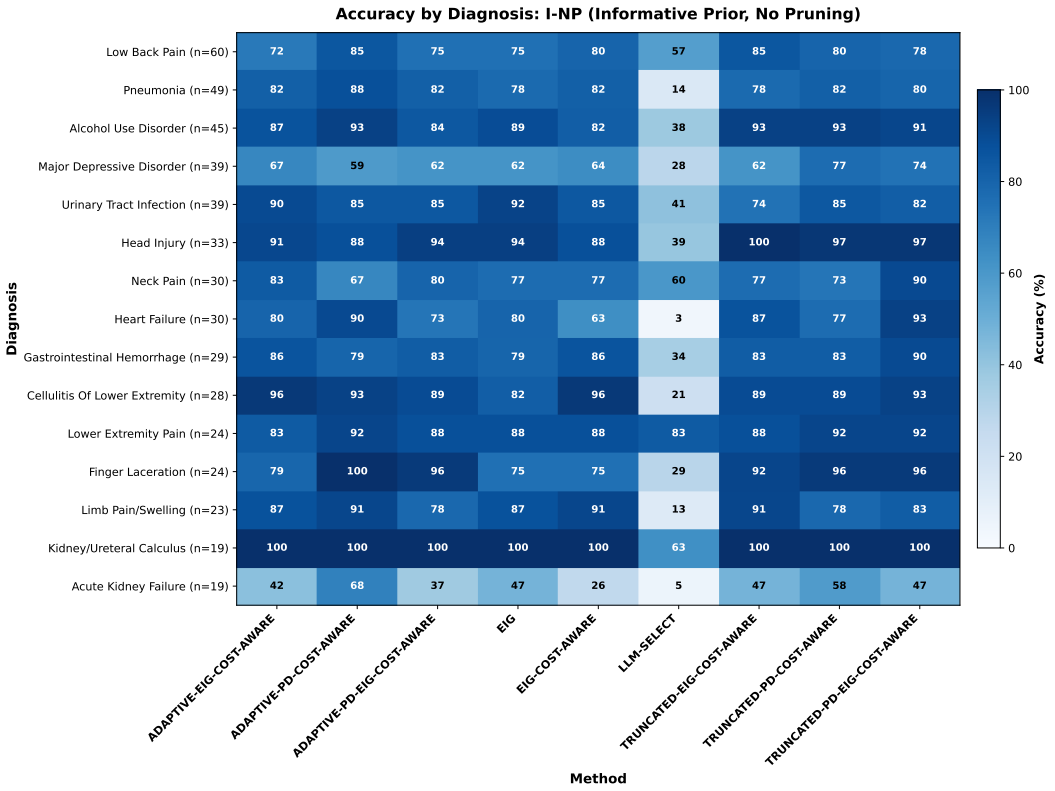


Figure 10: **Accuracy by diagnosis and method.** Each cell shows the percentage of correct predictions for a given diagnosis (row) and method (column). Diagnoses are sorted by frequency (top 15 most common). Darker blue indicates higher accuracy. LLM-Select clearly performs the worst. Major Depressive Disorder, one of the most frequent disease labels in the 1,000 random samples, is hard to diagnose across all methods.

Table 7: Discriminability analysis for I-NP (Informative Prior, No Pruning, $n = 1{,}000$). Cohen's d measures how well confidence separates correct from incorrect predictions. Higher values indicate better self-awareness.

Method	Conf (Correct)	Conf (Incorrect)	Gap	Cohen's d	Interpretation
EIG	0.927	0.707	+0.220	1.39	Very Large
Adaptive-EIG-CA	0.935	0.739	+0.196	1.38	Very Large
Truncated-EIG-CA	0.936	0.756	+0.180	1.37	Very Large
Adaptive-PD-CA	0.935	0.748	+0.187	1.36	Very Large
Adaptive-PD-EIG-CA	0.931	0.738	+0.193	1.30	Very Large
Truncated-PD-CA	0.938	0.773	+0.165	1.27	Very Large
Truncated-PD-EIG-CA	0.930	0.764	+0.166	1.25	Very Large
EIG-CA	0.926	0.741	+0.185	1.23	Very Large
LLM-Select	0.579	0.591	-0.012	-0.04	Inverted

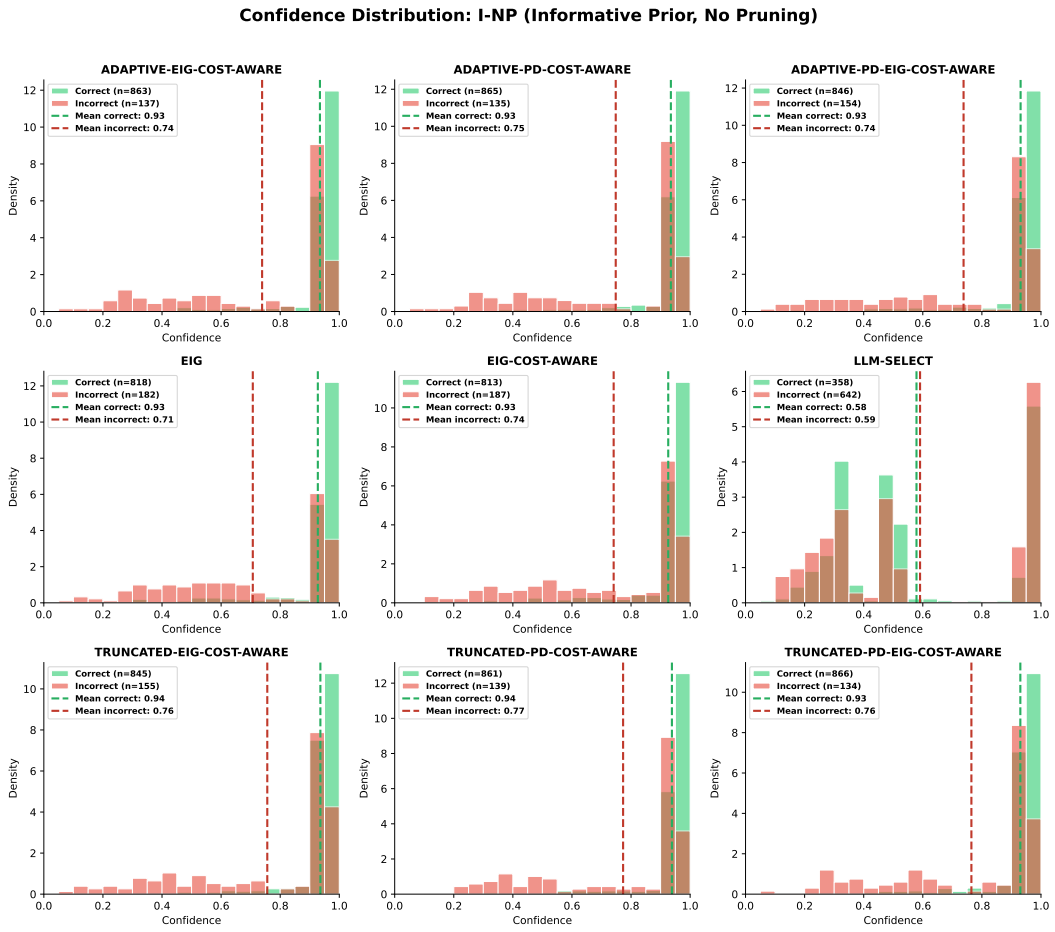


Figure 11: **Distribution of prediction confidence for correct (green) and incorrect (red) predictions across all methods.** Dashed vertical lines indicate mean confidence for each group. A well-calibrated method with good discriminability should show clear separation between the two distributions, with correct predictions concentrated at high confidence and incorrect predictions at lower confidence.

G.3 Not prune with uniform priors

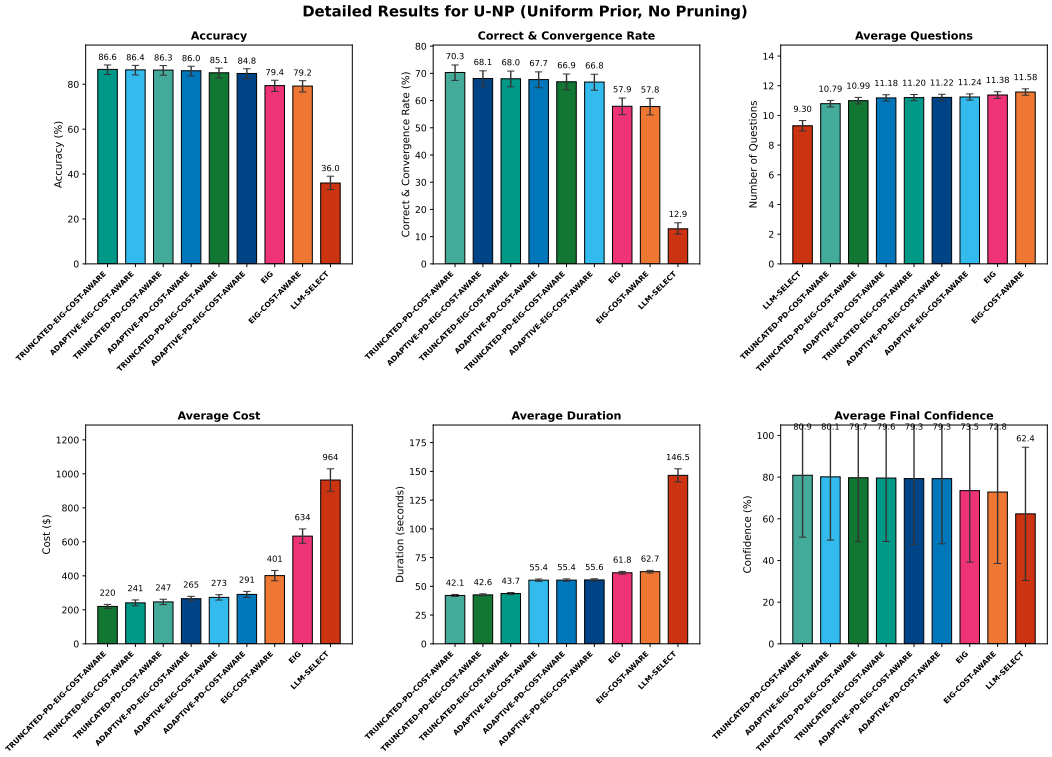


Figure 12: **Performance metrics for Uniform Prior with No Pruning (U-NP).** Starting from a flat distribution, this condition is the hardest stress-test. Truncated and adaptive strategies outperform standard EIG, demonstrating that active discrimination is crucial when initial uncertainty is maximal.

Table 8: Summary results for U-NP (Uniform Prior, No Pruning). Values show mean with 95% CI in brackets. The nominal sample is 1,000 visits; valid n is 994–1,000 depending on method.

Method	Acc (%)	C&C (%)	Questions	Cost ($)	Duration (s)
Truncated-EIG-CA	86.6 [84.4–88.6]	68.0 [65.0–70.8]	11.2 ± 0.2	241 ± 17	43.7 ± 0.9
Adaptive-EIG-CA	86.4 [84.1–88.4]	66.8 [63.8–69.7]	11.2 ± 0.2	273 ± 16	55.4 ± 1.0
Truncated-PD-CA	86.3 [84.0–88.3]	70.3 [67.4–73.1]	10.8 ± 0.2	247 ± 16	42.1 ± 0.9
Adaptive-PD-CA	86.0 [83.7–88.0]	67.7 [64.7–70.5]	11.2 ± 0.2	291 ± 18	55.4 ± 1.1
Truncated-PD-EIG-CA	85.1 [82.8–87.2]	66.9 [63.9–69.8]	11.0 ± 0.2	220 ± 12	42.6 ± 0.9
Adaptive-PD-EIG-CA	84.8 [82.4–86.9]	68.1 [65.1–70.9]	11.2 ± 0.2	265 ± 14	55.6 ± 1.0
EIG	79.4 [76.8–81.8]	57.9 [54.8–60.9]	11.4 ± 0.2	634 ± 42	61.8 ± 1.2
EIG-CA	79.2 [76.6–81.6]	57.8 [54.7–60.8]	11.6 ± 0.2	401 ± 30	62.7 ± 1.2
LLM-Select	36.0 [33.1–39.0]	12.9 [10.9–15.1]	9.3 ± 0.4	964 ± 66	146.5 ± 5.7

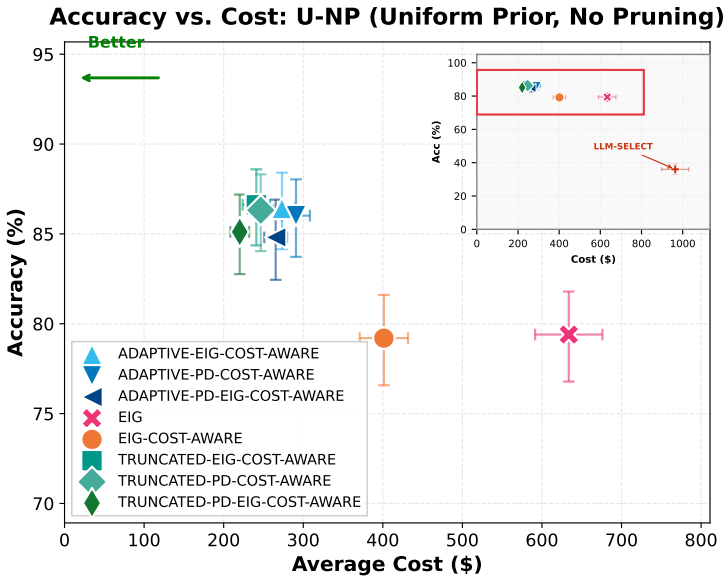


Figure 13: **Cost-Accuracy Pareto Frontier (U-NP).** Methods in the top-left corner represent the optimal trade-off (high accuracy, low cost). Adaptive and Truncated methods dominate the frontier, offering better diagnostic performance per dollar than standard baselines.

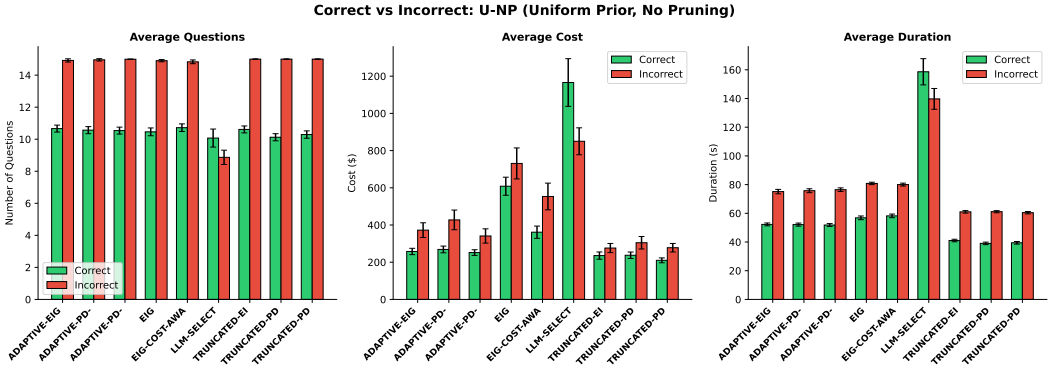


Figure 14: **Resource consumption for Correct vs. Incorrect diagnoses (U-NP).** Comparison of session length and cost for successful and failed sessions when starting with a uniform prior.

G.4 Prune with uniform priors

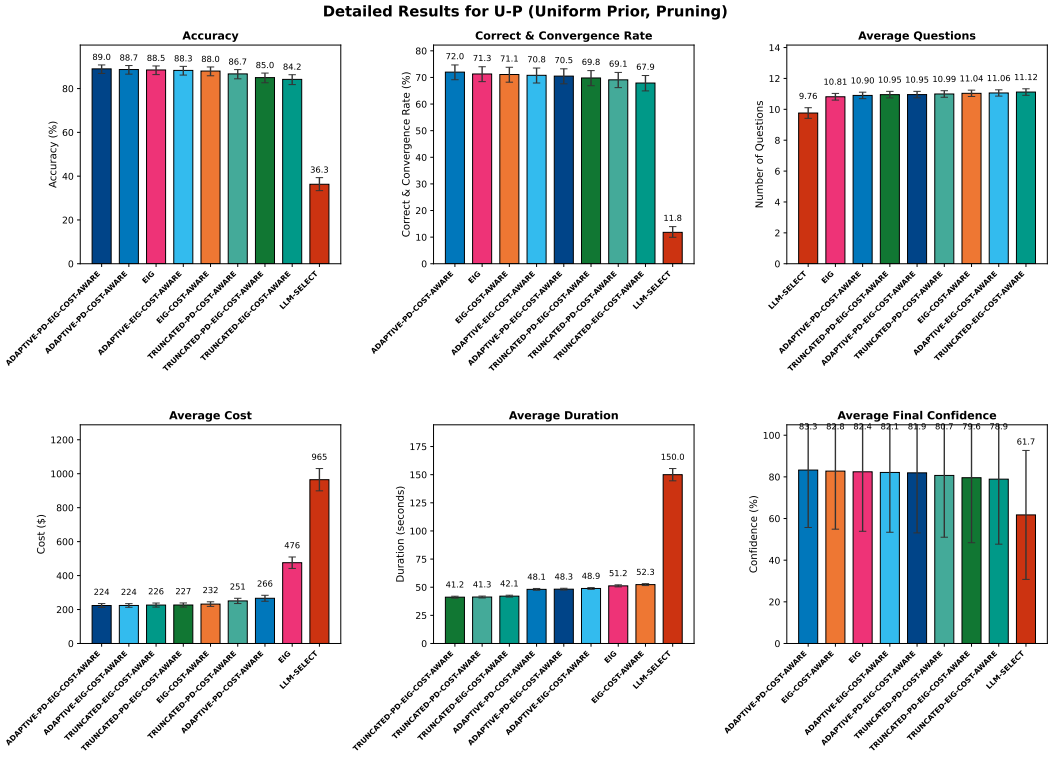


Figure 15: **Performance metrics for Uniform Prior with Pruning (U-P).** Pruning accelerates belief convergence. ADAPTIVE-PD-EIG-COST-AWARE achieves the highest accuracy in this set (89.0%), suggesting that eliminating low-probability candidates helps the strategy focus effectively.

Table 9: Summary results for U-P (Uniform Prior, Pruning, $n = 1{,}000$). Values show mean with 95% CI in brackets.

Method	Acc (%)	C&C (%)	Questions	Cost ($)	Duration (s)
Adaptive-PD-EIG-CA	89.0 [86.9–90.8]	70.5 [67.6–73.2]	11.0 ± 0.2	224 ± 12	48.3 ± 0.8
Adaptive-PD-CA	88.7 [86.6–90.5]	72.0 [69.1–74.7]	10.9 ± 0.2	266 ± 18	48.1 ± 0.8
EIG	88.5 [86.4–90.3]	71.3 [68.4–74.0]	10.8 ± 0.2	476 ± 34	51.2 ± 0.9
Adaptive-EIG-CA	88.3 [86.2–90.1]	70.8 [67.9–73.5]	11.1 ± 0.2	224 ± 11	48.9 ± 0.8
EIG-CA	88.0 [85.8–89.9]	71.1 [68.2–73.8]	11.0 ± 0.2	232 ± 13	52.3 ± 0.9
Truncated-PD-CA	86.7 [84.5–88.7]	69.1 [66.2–71.9]	11.0 ± 0.2	251 ± 16	41.3 ± 0.8
Truncated-PD-EIG-CA	85.0 [82.7–87.1]	69.8 [66.9–72.6]	10.9 ± 0.2	227 ± 12	41.2 ± 0.8
Truncated-EIG-CA	84.2 [81.8–86.3]	67.9 [64.9–70.7]	11.1 ± 0.2	226 ± 12	42.1 ± 0.8
LLM-Select	36.3 [33.4–39.3]	11.8 [9.9–13.9]	9.8 ± 0.3	965 ± 66	150.0 ± 5.5

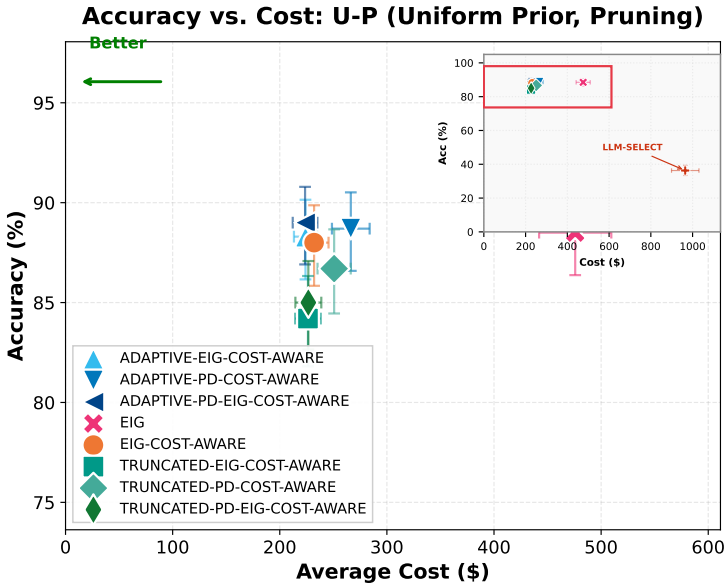


Figure 16: **Cost-Accuracy Pareto Frontier (U-P).** Pruning compresses the cost distribution, shifting most Bayesian methods toward lower costs while retaining high accuracy.

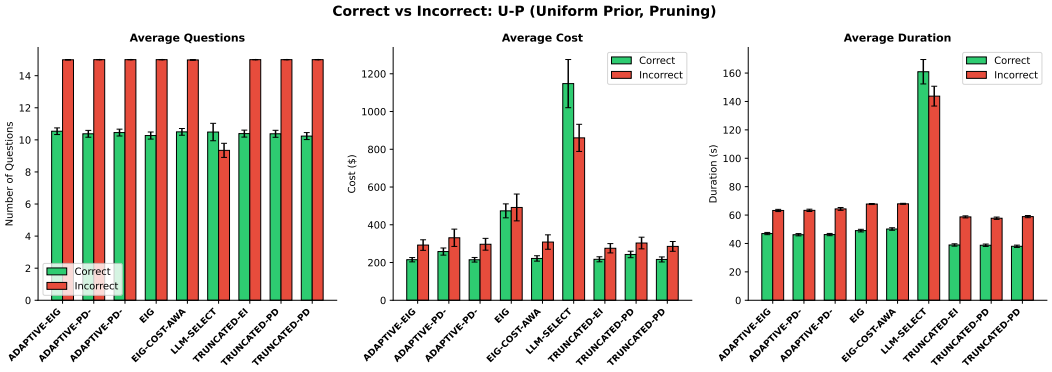


Figure 17: **Resource consumption for Correct vs. Incorrect diagnoses (U-P).** Comparison of session metrics for successful versus failed diagnostic runs under uniform priors with pruning.

G.5 Prune with informative priors

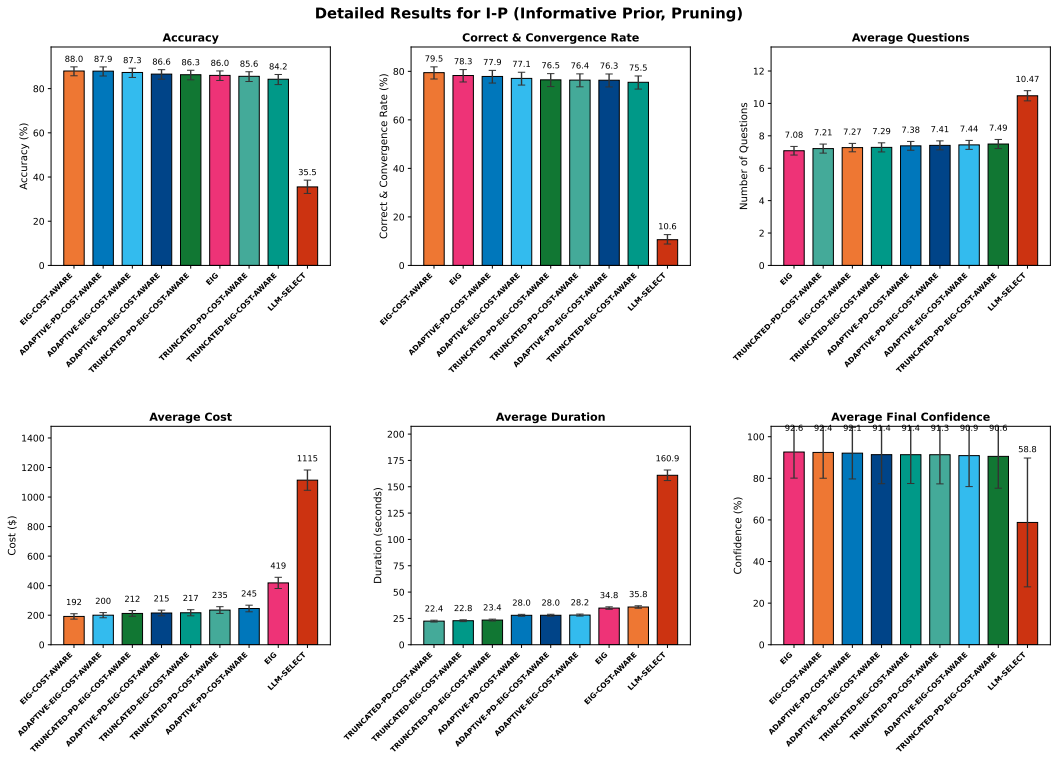


Figure 18: **Performance metrics for Informative Prior with Pruning (I-P).** This setting combines a strong starting prior with aggressive search space reduction. EIG-COST-AWARE achieves the highest accuracy (88.0%) in this condition, as pruning substantially amplifies its performance.

Table 10: Summary results for I-P (Informative Prior, Pruning). Values show mean with 95% CI in brackets. The nominal sample is 1,000 visits; valid n is 968–999 depending on method.

Method	Acc (%)	C&C (%)	Questions	Cost ($)	Duration (s)
EIG-CA	88.0 [85.8–89.8]	79.5 [76.8–81.9]	7.3 ± 0.3	192 ± 18	35.8 ± 1.2
Adaptive-PD-CA	87.9 [85.7–89.8]	77.9 [75.2–80.4]	7.4 ± 0.3	245 ± 22	28.0 ± 0.9
Adaptive-EIG-CA	87.3 [85.1–89.3]	77.1 [74.3–79.6]	7.4 ± 0.3	200 ± 17	28.2 ± 0.9
Adaptive-PD-EIG-CA	86.6 [84.3–88.6]	76.3 [73.6–78.9]	7.4 ± 0.3	215 ± 19	28.0 ± 0.9
Truncated-PD-EIG-CA	86.3 [84.0–88.3]	76.5 [73.7–79.1]	7.5 ± 0.3	212 ± 19	23.4 ± 0.9
EIG	86.0 [83.7–88.0]	78.3 [75.6–80.7]	7.1 ± 0.3	419 ± 38	34.8 ± 1.2
Truncated-PD-CA	85.6 [83.2–87.6]	76.4 [73.6–79.0]	7.2 ± 0.3	235 ± 23	22.4 ± 0.9
Truncated-EIG-CA	84.2 [81.8–86.4]	75.5 [72.7–78.1]	7.3 ± 0.3	217 ± 21	22.8 ± 0.9
LLM-Select	35.5 [32.6–38.6]	10.6 [8.8–12.7]	10.5 ± 0.3	1115 ± 69	160.9 ± 5.0

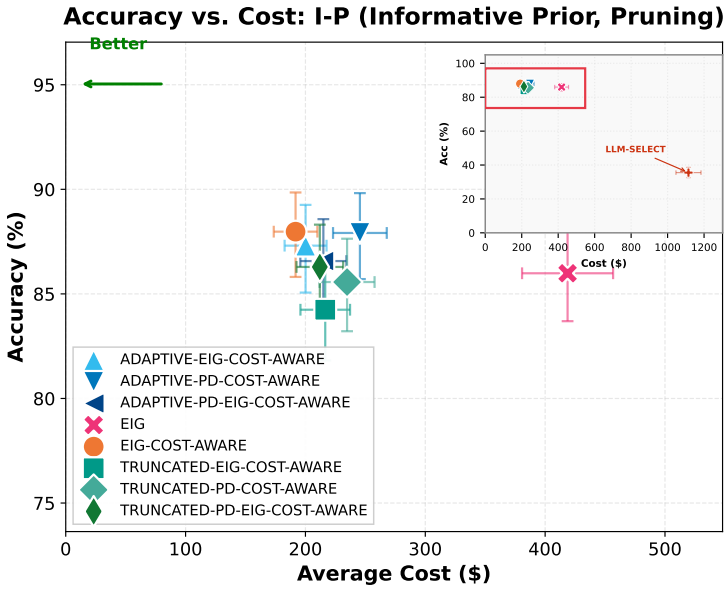


Figure 19: **Cost-Accuracy Pareto Frontier (I-P).** Pruning substantially improves the performance of EIG-COST-AWARE, which achieves the highest accuracy (88.0%) in this condition. Adaptive methods combined with pruning remain competitive.

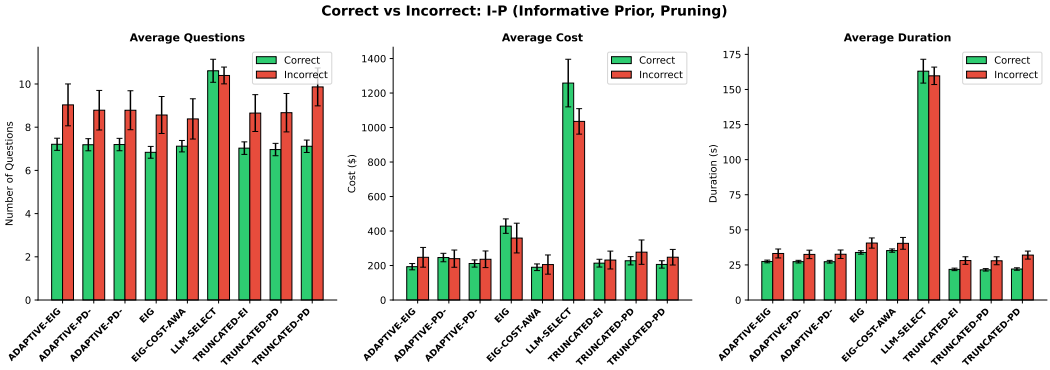


Figure 20: **Resource consumption for Correct vs. Incorrect diagnoses (I-P).** Comparison of cost and turns for successful and failed sessions when using both informative priors and pruning.

H Additional robustness studies

We include additional robustness studies to test whether the main conclusions depend on the particular model or selection mechanism used in the experiment (Section 5). Unless otherwise noted, these studies use the informative-prior, no-pruning setting (I-NP).

H.1 Qwen-4B

We replace MedGemma-4B with the domain-general `Qwen/Qwen3-4B-Instruct-2507` model and evaluate the same nine methods on 200 patients. Performance drops substantially relative to MedGemma-4B. EIG-Cost-Aware reaches 71.0% accuracy, while the best adaptive-family method, Adaptive-PD-Cost-Aware, reaches 73.0%. Thus, some adaptive-family methods remain competitive, but the adaptive-family advantage is not uniform under a non-medical model. This experiment was completed within 7 hours on an internal compute cluster equipped with NVIDIA RTX 6000 Ada Generation GPUs (48 GB VRAM).

Table 11: Qwen-4B robustness study under I-NP ($n = 200$). Values show mean with 95% CI in brackets.

Method	Acc (%)	C&C (%)	Questions	Cost ($)	Duration (s)
Adaptive-PD-CA	73.0 [66.5–78.7]	72.5 [65.9–78.2]	5.0 ± 0.4	360 ± 215	46.7 ± 4.6
Truncated-PD-CA	71.0 [64.4–76.8]	70.0 [63.3–75.9]	4.7 ± 0.4	276 ± 81	23.3 ± 2.4
EIG-CA	71.0 [64.4–76.8]	70.0 [63.3–75.9]	5.0 ± 0.4	157 ± 47	58.9 ± 5.1
Adaptive-EIG-CA	70.0 [63.3–75.9]	67.5 [60.7–73.6]	4.8 ± 0.4	264 ± 147	45.5 ± 4.5
Truncated-EIG-CA	69.5 [62.8–75.5]	67.0 [60.2–73.1]	5.0 ± 0.4	494 ± 316	23.3 ± 2.3
EIG	68.5 [61.8–74.5]	66.0 [59.2–72.2]	4.7 ± 0.4	681 ± 256	55.5 ± 5.3
Adaptive-PD-EIG-CA	67.5 [60.7–73.6]	65.0 [58.2–71.3]	5.0 ± 0.5	322 ± 199	45.1 ± 4.6
Truncated-PD-EIG-CA	67.5 [60.7–73.6]	67.0 [60.2–73.1]	4.9 ± 0.4	456 ± 216	23.6 ± 2.2
LLM-Select	22.0 [16.8–28.2]	7.5 [4.6–12.0]	8.3 ± 0.9	1390 ± 276	325.4 ± 36.8

H.2 Qwen-30B

We also evaluate a larger domain-general model, `Qwen/Qwen3-30B-A3B-Instruct-2507`, on 200 patients using a subset of cost-aware methods. Qwen-30B improves over Qwen-4B but remains below the MedGemma-4B main results. Adaptive-PD-Cost-Aware slightly outperforms EIG-Cost-Aware (81.0% vs. 80.5%), while Truncated-PD-EIG-Cost-Aware falls below EIG-Cost-Aware, again showing that adaptive-family gains depend on the model–task pairing. This experiment was completed within 1 hour on a Google Cloud Platform (GCP) instance equipped with NVIDIA A100-SXM4-80GB GPUs.

Table 12: Qwen-30B robustness study under I-NP ($n = 200$). Values show mean with 95% CI in brackets.

Method	Acc (%)	C&C (%)	Questions	Cost ($)	Duration (s)
Adaptive-PD-CA	81.0 [75.0–85.8]	80.5 [74.5–85.4]	4.7 ± 0.5	148 ± 37	159.6 ± 17.4
EIG-CA	80.5 [74.5–85.4]	79.0 [72.8–84.1]	4.8 ± 0.4	163 ± 63	230.4 ± 21.0
Truncated-PD-CA	80.0 [73.9–85.0]	79.5 [73.4–84.5]	4.7 ± 0.5	182 ± 55	108.1 ± 11.4
Truncated-PD-EIG-CA	76.0 [69.6–81.4]	74.5 [68.0–80.0]	4.6 ± 0.5	173 ± 57	109.3 ± 11.4

H.3 MedGemma-27B

We replace MedGemma-4B with `google/medgemma-27b-text-it` and evaluate 100 patients due to the larger model size. This experiment was completed within 4 hours on a Google Cloud Platform (GCP) instance equipped with NVIDIA A100-SXM4-80GB GPUs.

Score-based methods reach 99–100% accuracy, while LLM-Select remains much worse at 35.0%. These near-ceiling score-based results are striking but should be interpreted

cautiously: the MedGemma documentation references related MIMIC resources, so possible pretraining exposure or benchmark contamination cannot be ruled out.

Table 13: MedGemma-27B robustness study under I-NP ($n = 100$). Values show mean with 95% CI in brackets.

Method	Acc (%)	C&C (%)	Questions	Cost ($)	Duration (s)
Adaptive-PD-CA	100.0 [96.3–100.0]	97.0 [91.5–99.0]	7.8 ± 0.5	186 ± 49	1036.4 ± 83.6
Adaptive-PD-EIG-CA	100.0 [96.3–100.0]	100.0 [96.3–100.0]	8.0 ± 0.4	151 ± 17	1020.8 ± 75.8
Adaptive-EIG-CA	99.0 [94.6–99.8]	96.0 [90.2–98.4]	7.8 ± 0.6	197 ± 57	1035.1 ± 77.4
EIG	99.0 [94.6–99.8]	95.0 [88.8–97.8]	7.4 ± 0.5	986 ± 221	1289.9 ± 102.5
EIG-CA	99.0 [94.6–99.8]	96.0 [90.2–98.4]	8.0 ± 0.6	153 ± 34	1385.0 ± 100.9
Truncated-EIG-CA	99.0 [94.6–99.8]	95.0 [88.8–97.8]	8.1 ± 0.5	188 ± 47	845.9 ± 49.8
Truncated-PD-CA	99.0 [94.6–99.8]	97.0 [91.5–99.0]	8.0 ± 0.6	289 ± 159	810.6 ± 52.6
Truncated-PD-EIG-CA	99.0 [94.6–99.8]	93.0 [86.3–96.6]	8.2 ± 0.6	180 ± 45	825.0 ± 60.9
LLM-Select	35.0 [26.4–44.7]	0.0 [0.0–3.7]	15.0 ± 0.0	1596 ± 228	1120.4 ± 17.5

H.4 Random-EIG selection

Random-EIG isolates the contribution of EIG scoring. It uses the same candidate-generation and Bayesian-update pipeline as EIG, but replaces EIG maximization with random selection from the valid candidate questions/tests at each turn. This experiment was completed within 2 hours on an internal compute cluster equipped with NVIDIA RTX 6000 Ada Generation GPUs (48 GB VRAM).

On 200 patients, random selection reduces accuracy from 78.0% to 64.0%, reduces C&C from 69.0% to 36.5%, and increases average cost from $535 to $1,316. This shows that the LLM-generated candidate pool alone is insufficient: explicit scoring is needed to choose useful and cost-efficient queries.

Table 14: Random-EIG selection ablation under I-NP ($n = 200$). Values show mean with 95% CI in brackets.

Method	Acc (%)	C&C (%)	Questions	Cost ($)	Duration (s)
EIG	78.0 [71.8–83.2]	69.0 [62.3–75.0]	7.4 ± 0.6	535 ± 94	50.4 ± 4.9
Random-EIG	64.0 [57.1–70.3]	36.5 [30.1–43.4]	11.3 ± 0.6	1316 ± 164	34.2 ± 2.0
LLM-Select	32.0 [25.9–38.8]	5.5 [3.1–9.6]	11.3 ± 0.7	1288 ± 163	246.1 ± 21.1

H.5 Split Questioner–Oracle models

We also evaluate a split Questioner–Oracle/likelihood setting to address concerns that using the same model throughout can make the evaluation overly self-consistent. Here `google/medgemma-4b-it` proposes candidate questions/tests, while `google/medgemma-27b-text-it` is served separately through vLLM to provide Oracle answers and the likelihood estimates used by score-based methods. Because we require serving both models and using the larger 27B model in the likelihood-estimation loop, we run the study on 100 patients under I-NP. This experiment was completed within 8 hours on a Google Cloud Platform (GCP) instance equipped with NVIDIA A100-SXM4-80GB GPUs.

The best adaptive-family method is Truncated-PD-Cost-Aware at 87.0% accuracy, compared with 84.0% for EIG and 82.0% for EIG-Cost-Aware. Truncated-EIG-Cost-Aware also improves over both baselines at 85.0%. However, the adaptive-family advantage is not uniform: the remaining adaptive/truncated variants range from 81.0% to 83.0% accuracy. Adaptive/truncated cost-aware variants, averaging $207–$294 per session, reduce average cost substantially relative to both EIG ($1,130) and EIG-Cost-Aware ($501).

Table 15: Split Questioner–Oracle/likelihood robustness study under I-NP ($n = 100$). Values show mean with 95% CI in brackets.

Method	Acc (%)	C&C (%)	Questions	Cost ($)	Duration (s)
Truncated-PD-CA	87.0 [79.0–92.2]	79.0 [70.0–85.8]	6.8 ± 0.8	268 ± 78	272.0 ± 30.4
Truncated-EIG-CA	85.0 [76.7–90.7]	79.0 [70.0–85.8]	6.6 ± 0.8	228 ± 71	265.9 ± 29.1
EIG	84.0 [75.6–89.9]	72.0 [62.5–79.9]	7.1 ± 0.8	1130 ± 249	586.3 ± 63.9
Adaptive-EIG-CA	83.0 [74.5–89.1]	77.0 [67.8–84.2]	6.5 ± 0.8	284 ± 107	364.8 ± 45.8
Adaptive-PD-EIG-CA	82.0 [73.3–88.3]	75.0 [65.7–82.5]	6.6 ± 0.8	239 ± 61	354.6 ± 46.7
Truncated-PD-EIG-CA	82.0 [73.3–88.3]	79.0 [70.0–85.8]	6.3 ± 0.7	207 ± 56	253.0 ± 28.7
EIG-CA	82.0 [73.3–88.3]	72.0 [62.5–79.9]	7.0 ± 0.8	501 ± 165	583.2 ± 63.8
Adaptive-PD-CA	81.0 [72.2–87.5]	75.0 [65.7–82.5]	6.5 ± 0.8	294 ± 100	360.4 ± 44.6
LLM-Select	35.0 [26.4–44.7]	9.0 [4.8–16.2]	8.6 ± 1.0	1056 ± 204	86.7 ± 9.8