WHICH AI AGENT ACTUALLY WINS? ACTIVE QUERY GENERATION FOR HIGH-CONFIDENCE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Choosing an artificial intelligence (AI) agent for deployment can depend on a small subset of queries that distinguish otherwise similar candidates. These diagnostic queries may be rare and unknown in advance. Evaluating agents on fixed benchmarks or randomly sampled queries can therefore spend much of the evaluation budget on cases that do little to distinguish them. To address this problem, we formulate a framework for selecting the best agent under a limited evaluation budget, by adaptively choosing which queries to evaluate. Within this framework, we propose Diagnostic Agent Selection via Exploration (DASE), which directs evaluation toward distinguishing the most competitive candidates. At each round, DASE identifies the current leader and a challenger, then generates queries likely to distinguish them using previous evaluations. The generator can be adapted to the application, for example by using a large language model conditioned on the evaluation history. DASE selects one of these queries, evaluates all agents on it, and updates its estimates. Theoretically, we provide lower bounds on the number of diagnostic queries selected and the probability of selecting the best agent. Empirically, we evaluate DASE in three settings: simulated agent selection, mathematical reasoning, and clinical decision-making. Across these settings, DASE generally achieves higher best-agent selection accuracy than the tested baselines, particularly at larger evaluation budgets.

1 INTRODUCTION

Deploying an artificial intelligence (AI) agent in a high-stakes setting requires reliable evidence that it is preferable to available alternatives, often from costly empirical evaluation. For example, a robotics team choosing among control policies must stage physical scenes and run the robot through each. Similarly, a software team comparing coding agent configurations must set up sandboxed environments and execute full test suites. Such trials can be expensive in time, computation, and operational effort. With only a limited budget, the practitioner must decide, round by round, which query to run next.

A natural approach is to test each candidate on queries it would face in deployment, such as a standard pick-and-place suite or the next batch of bug reports, and estimate performance by averaging. This can be inefficient when most queries provide little information for distinguishing the candidates. A routine pick-and-place motion may be solved by several control policies, just as a simple bug fix may be handled by several coding agents. Differences may emerge only on a small, often unknown subset of queries. For example, an edge case may expose a difference in race-condition handling, or a particular configuration may reveal a planner failure that routine cases do not. We refer to queries that meaningfully separate candidate agents as *diagnostic queries*. When such queries are rare, fixed benchmarks or random sampling may use much of the budget on weakly informative cases.

This focus on diagnostic queries connects our problem to prior work on adaptive evaluation. Best-arm identification seeks the best option under a limited sampling budget (Kaufmann et al., 2016), and top-two sampling rules focus effort on a leader and a challenger (Russo, 2016). Our setting also requires choosing the shared query on which candidates are compared. Active testing and active model comparison select test instances to estimate or compare the risks of fixed models (Sawade et al., 2012; Kossen et al., 2021), and recent work applies active selection to large language mod-

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.

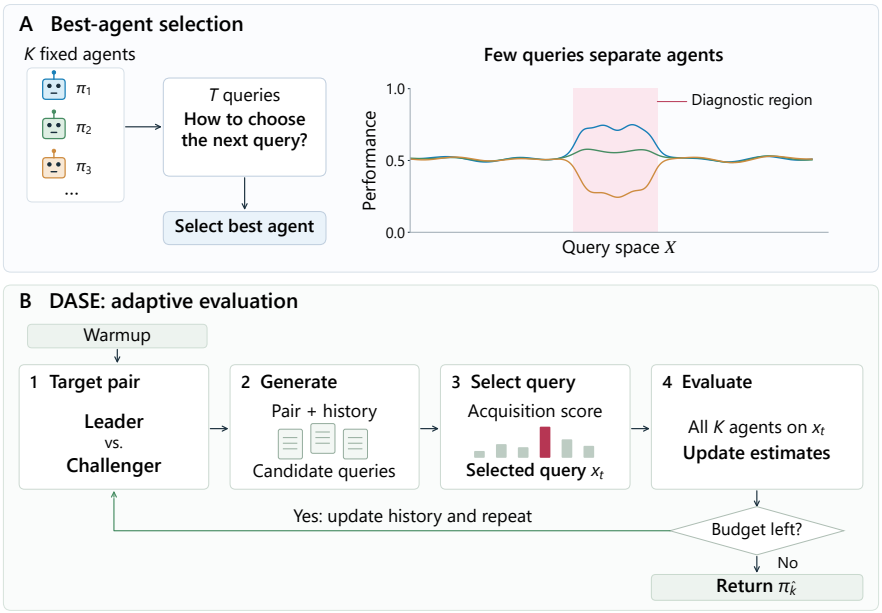

Figure 1: **Best-agent selection and DASE. (A)** Select among K fixed agents using T queries. Schematic curves illustrate a rare diagnostic region that distinguishes the agents. **(B)** After warmup, DASE targets a leader–challenger pair, generates and selects queries, and evaluates all agents, updating estimates until the budget is exhausted.

els (LLMs) on annotated benchmarks (Durmazkeser et al., 2025; Tolochinsky et al., 2026). We also allow queries to be generated outside the initial pool. Dynamic benchmarking and automated benchmark construction create challenging examples to expose model failures or build harder benchmarks (Kiela et al., 2021; Li et al., 2025).

In this work, we aim to identify the best agent among a fixed set of candidates under a limited evaluation budget, by adaptively generating and selecting queries. We propose Diagnostic Agent Selection via Exploration (DASE), an adaptive algorithm for this task. Given K fixed candidate agents and a budget of T rounds, each round selects one query and observes every candidate's outcome. The goal is to identify the best agent with probability at least $1 - \delta$, where $\delta \in (0, 1)$ is a target error probability. DASE directs evaluation toward a pair selected from the current estimates. It tracks the *leader*, the candidate with the highest current estimated value, and the *challenger*, the non-leading candidate with the highest uncertainty-adjusted value estimate. Each active round then targets this pair (Figure 1).

The key component is a generator that seeks new diagnostic queries for the current leader–challenger pair. Rather than relying only on queries already available in a fixed pool, the generator uses the evaluation history to identify promising regions and proposes related candidates. For example, it may generate variants of previously observed queries on which the leader and challenger differed substantially. As evidence accumulates, the leader-challenger pair can change, and generation is redirected accordingly. The generator is domain-dependent: it can perturb numerical query representations, or use an LLM to propose queries conditioned on the evaluation history.

We make three contributions. First, we formalize best-agent selection under a limited evaluation budget and define pairwise diagnostic regions. Second, we introduce DASE, which generates and selects queries to distinguish an adaptively chosen leader–challenger pair. Third, we provide theoretical and empirical analyses of this strategy. We derive lower bounds on the number of diagnostic queries selected and the probability of selecting the best agent. We evaluate DASE in three settings: a synthetic sparse-signal problem, a comparison of four LLM prompting strategies and a constructed comparator on the MATH benchmark (Hendrycks et al., 2021), and clinical decision-making by AI agents using records from real hospital admissions (Hager et al., 2024). The generator samples from Gaussian distributions centered at evaluated queries in the synthetic study and retrieves queries from

a separate pool in each real-data study. Across the three studies, DASE generally achieves higher best-agent selection accuracy than the tested random-sampling and query-by-committee baselines, particularly at larger evaluation budgets. These results show that query generation can improve best-agent selection under a limited evaluation budget.

2 Problem Setup

2.1 Setting

Let $\Pi = \{\pi_1, \ldots, \pi_K\}$ be a set of $K \geq 2$ fixed candidate AI agents, indexed by $[K] := \{1, \ldots, K\}$. For convenience, we refer to agent π_k by its index k. A *query* $x \in \mathcal{X}$ specifies a complete evaluation instance, such as a mathematical problem or a clinical case, and $\mathcal{X}$ denotes the query space. Each agent maps a query to an action or output. Evaluating this response produces a numerical outcome $Y_k(x)$, with higher outcomes indicating better performance. Its finite mean $\mu_k(x) := \mathbb{E}[Y_k(x)]$ is the agent's *conditional value*, or expected performance on query x. Here $\mathbb{E}$ denotes expectation over evaluation randomness at a fixed query.

Let $\mathcal{D}$ denote the *deployment distribution*, the distribution of queries on which performance is to be assessed. Assume each μ_k is integrable under $\mathcal{D}$. The *deployment value* $V_k := \mathbb{E}_{X\sim\mathcal{D}}[\mu_k(X)]$ is agent k's expected performance for a query X drawn from $\mathcal{D}$. Assume there is a unique best agent, with index $k^* := \arg\max_{k\in[K]} V_k$. For $k \neq k^*$, the *value gap* $\Delta_k := V_{k^*} - V_k$ measures its shortfall relative to the best agent. The *minimum value gap* is $\Delta_{\min} := \min_{k\neq k^*} \Delta_k > 0$.

We consider an evaluation budget of T rounds. At each round t, a query $x_t \in \mathcal{X}$ is selected and all K agents are evaluated, yielding outcomes $\{Y_k(x_t)\}_{k=1}^K$. The goal is to return an agent index $\hat{k} \in [K]$ within T rounds such that $\mathbb{P}(\hat{k} = k^*) \geq 1 - \delta$, where $\mathbb{P}$ denotes probability and $\delta \in (0, 1)$ is the target error probability.

A simple baseline samples queries from $\mathcal{D}$ and estimates each V_k by averaging the observed outcomes. However, the difference $|\mu_i(x) - \mu_j(x)|$ between two agents' conditional values may be appreciable only on a small subset of the query space. If this subset has low probability under $\mathcal{D}$, random sampling encounters few queries that help distinguish the agents.

2.2 Diagnostic regions

For distinct agents i, j, define the *conditional value gap* $D_{ij}(x) := \mu_i(x) - \mu_j(x)$. For a threshold $u > 0$, the *diagnostic region* contains queries on which the gap magnitude is at least u: $\mathcal{X}_{ij}(u) := \{x \in \mathcal{X} : |D_{ij}(x)| \geq u\}$. The probability that a deployment query falls in this region is $\rho_{ij}(u) := \mathbb{P}_{X\sim\mathcal{D}}\big(X \in \mathcal{X}_{ij}(u)\big)$.

The symbol u denotes a general threshold. For each pair, fix a reference threshold $\gamma_{ij} = \gamma_{ji} > 0$ before evaluation and write $\mathcal{X}_{ij} := \mathcal{X}_{ij}(\gamma_{ij})$. These diagnostic regions are unknown. We use *diagnostic sparsity* to describe the regime $\rho_{ij}(\gamma_{ij}) \ll 1$, in which diagnostic queries are rare under $\mathcal{D}$. This motivates generating queries targeted to a pair of agents.

2.3 Query generation

Let $\mathcal{H}$ denote the *evaluation history*, which records the queries already evaluated and the corresponding outcomes of all K agents. A *query generator* $\mathcal{G}$ is a procedure that takes an agent pair (i, j) and history $\mathcal{H}$ as input and returns N candidate queries, $\mathcal{C}_{\text{gen}} = \mathcal{G}(i, j, \mathcal{H}) \subseteq \mathcal{X}$, where N is the number of candidates proposed per call. These are queries proposed for evaluation, using the available evidence to identify possible differences between the pair. Section 3 specifies how DASE chooses the pair, generates candidates, and selects a query.

3 ALGORITHM

3.1 OVERVIEW

DASE begins with a warmup phase that samples queries from $\mathcal{D}$ to initialize the estimator and the evaluation history $\mathcal{H}$. The active phase then selects a target pair, generates candidates, chooses one query for evaluation, and updates the history.

At each active round, DASE identifies a *critical pair*: a leader $\hat{\imath}$ and a challenger $\hat{\jmath}$. The leader has the highest estimated deployment value. Among the remaining agents, the challenger maximizes an upper confidence bound (UCB)-style score that adds an uncertainty bonus to each agent's estimated deployment value (Auer et al., 2002). DASE generates and selects queries to distinguish this pair.

3.2 EVALUATION PROCEDURE

Warmup. DASE draws T_0 queries, with $1 \leq T_0 < T$, from $\mathcal{D}$ and evaluates all K agents on each. These observations form the initial evaluation history used by the estimator and generator.

Leader and challenger selection. At active round t, the estimator $\mathcal{E}$ uses the evaluation history $\mathcal{H}$ to compute a conditional-value estimate $\hat{\mu}_k(\cdot)$ and an uncertainty proxy $\hat{\sigma}_k(\cdot)$ for each agent k. The proxy measures estimated uncertainty in the prediction at a query. DASE averages both quantities over the deployment distribution: $\hat{V}_k = \mathbb{E}_{x\sim\mathcal{D}}[\hat{\mu}_k(x)]$ and $\hat{S}_k = \mathbb{E}_{x\sim\mathcal{D}}[\hat{\sigma}_k(x)]$. Here $\hat{V}_k$ estimates deployment value, and $\hat{S}_k$ is the average uncertainty proxy. The leader is $\hat{\imath} = \arg\max_{k\in[K]} \hat{V}_k$. Among the remaining agents, DASE selects the challenger with the highest UCB-style score $\hat{\jmath} = \arg\max_{k\neq\hat{\imath}}[\hat{V}_k + \beta_t \hat{S}_k]$, where $\beta_t > 0$ controls the uncertainty bonus. The critical pair may change as additional evidence is collected.

Candidate generation. DASE calls the generator defined in Section 2.3 with the current leader–challenger pair $(\hat{\imath}, \hat{\jmath})$ and history $\mathcal{H}$ to produce the candidate set $\mathcal{C}_{\text{gen}} = \mathcal{G}(\hat{\imath}, \hat{\jmath}, \mathcal{H})$, targeting the diagnostic region $\mathcal{X}_{\hat{\imath}\hat{\jmath}}$.

DASE takes a user-specified query generator $\mathcal{G}$. The aim is to propose queries in $\mathcal{X}_{\hat{\imath}\hat{\jmath}}$ more frequently than sampling from $\mathcal{D}$.

Query acquisition. DASE can supplement generated candidates with a fixed candidate pool $\mathcal{C}_{\text{pool}}$, a collection of queries available before evaluation begins. It forms $\mathcal{C} = \mathcal{C}_{\text{gen}} \cup \mathcal{C}_{\text{pool}}$, with $\mathcal{C}_{\text{pool}} = \emptyset$ when no pool is used, and selects the query maximizing the acquisition score

$$\mathcal{A}_t(x) = |\hat{\mu}_{\hat{\imath}}(x) - \hat{\mu}_{\hat{\jmath}}(x)| + \beta_t \left(\hat{\sigma}_{\hat{\imath}}(x) + \hat{\sigma}_{\hat{\jmath}}(x)\right). \tag{1}$$

The first term favors queries with larger estimated conditional-value gaps between the leader and challenger. The second favors queries with larger uncertainty proxies. DASE evaluates all K agents on the selected query and appends the query and outcomes to $\mathcal{H}$.

After T rounds, DASE updates the estimates using the complete history and returns $\hat{k} = \arg\max_{k\in[K]} \hat{V}_k$. Algorithm 1 summarizes DASE.

3.3 PRACTICAL ESTIMATOR

DASE can use any estimator that supplies conditional-value estimates and uncertainty proxies. Our implementation uses a Nadaraya–Watson (NW) estimator (Nadaraya, 1964; Watson, 1964), which predicts a query's conditional value by averaging observed outcomes with larger weights for nearby queries. We use the Gaussian kernel $K_h(x, x') = \exp(-\|x - x'\|^2/2h^2)$, where $h > 0$ is the bandwidth and $\|x - x'\|$ is Euclidean distance between numerical query representations. The normalized weights are $A_{t,s}(x) = K_h(x, x_s)/\sum_{r<t} K_h(x, x_r)$.

At active round t, the estimates use observations $s < t$; the final estimates use all T observations. We include the round index t explicitly below and define $Q_t(x) := \sum_{s<t} A_{t,s}(x)^2$, which measures

Algorithm 1 DASE for best-agent selection

Require: Candidate agents $\{\pi_k\}_{k=1}^K$; budget T; warmup budget T_0; deployment distribution $\mathcal{D}$; generator $\mathcal{G}$ producing N candidates per round; candidate pool $\mathcal{C}_{\text{pool}}$; estimator $\mathcal{E}$ returning $\{(\hat{\mu}_k, \hat{\sigma}_k)\}_{k=1}^K$; uncertainty-bonus coefficient β_t.

Ensure: Selected agent index $\hat{k}$.

Draw $x_1, \ldots, x_{T_0} \sim \mathcal{D}$ and evaluate all agents.
Initialize $\mathcal{H}$ with $(x_s, \{Y_k(x_s)\}_{k=1}^K)$ for $s = 1, \ldots, T_0$.
for $t = T_0 + 1, \ldots, T$ **do**
    Compute $\{(\hat{\mu}_k, \hat{\sigma}_k)\}_{k=1}^K \leftarrow \mathcal{E}(\mathcal{H})$.
    $\hat{V}_k \leftarrow \mathbb{E}_{x\sim\mathcal{D}}[\hat{\mu}_k(x)]$ and $\hat{S}_k \leftarrow \mathbb{E}_{x\sim\mathcal{D}}[\hat{\sigma}_k(x)]$ for all k.
    $\hat{\imath} \leftarrow \arg\max_k \hat{V}_k$.
    $\hat{\jmath} \leftarrow \arg\max_{k\neq\hat{\imath}}[\hat{V}_k + \beta_t \hat{S}_k]$.
    $\mathcal{C}_{\text{gen}} \leftarrow \mathcal{G}(\hat{\imath}, \hat{\jmath}, \mathcal{H})$.
    $\mathcal{C} \leftarrow \mathcal{C}_{\text{gen}} \cup \mathcal{C}_{\text{pool}}$.
    Select $x_t \leftarrow \arg\max_{x\in\mathcal{C}} \mathcal{A}_t(x)$, with $\mathcal{A}_t$ defined in equation 1.
    Observe $\{Y_k(x_t)\}_{k=1}^K$ and append $(x_t, \{Y_k(x_t)\}_{k=1}^K)$ to $\mathcal{H}$.
end for
Recompute $\{(\hat{\mu}_k, \hat{\sigma}_k)\}_{k=1}^K \leftarrow \mathcal{E}(\mathcal{H})$ and $\hat{V}_k \leftarrow \mathbb{E}_{x\sim\mathcal{D}}[\hat{\mu}_k(x)]$.
return $\hat{k} \leftarrow \arg\max_k \hat{V}_k$.

concentration of the normalized weights. The conditional-value estimate and uncertainty proxy are

$$\hat{\mu}_{k,t}(x) = \sum_{s<t} A_{t,s}(x) Y_k(x_s), \quad \hat{\sigma}_{k,t}(x) = \frac{\sqrt{Q_t(x)}}{\sigma} \left[\sum_{s<t} A_{t,s}(x) \left(Y_k(x_s) - \hat{\mu}_{k,t}(x)\right)^2\right]^{1/2},$$

Here $\sigma > 0$ is a noise-scale parameter for the uncertainty proxy.

4 Theoretical Analysis

We first analyze how many diagnostic queries DASE selects and when this number exceeds that obtained by random sampling from $\mathcal{D}$ for a fixed agent pair. We then give conditions on estimation error that ensure the best agent is selected with probability at least $1 - \delta$.

Let $B := T - T_0 \geq 1$ be the number of active rounds, indexed here by $t = 1, \ldots, B$, and let (I_t, J_t) denote the leader–challenger pair. The filtration $\{\mathcal{F}_t\}_{t=0}^B$ represents the information available as evaluation proceeds. The information in $\mathcal{F}_{t-1}$ includes the history and target pair before round t's candidate generation; $\mathcal{F}_t$ also includes that round's candidates, selected query, and outcomes. The target pair is therefore predictable, meaning that it is determined by $\mathcal{F}_{t-1}$. Candidates may depend on this history and on each other. Write $\mathcal{C}_t = \mathcal{C}_{\text{gen},t} \cup \mathcal{C}_{\text{pool}}$ for the candidate set and x_t for the selected query in active round t.

4.1 Diagnostic-query acquisition

Fix symmetric thresholds $0 < u_{ij} \leq \gamma_{ij}$ before evaluation. We use these two thresholds to distinguish candidate generation from query selection. For a target pair (i, j), a candidate set offers a diagnostic opportunity if it contains a query whose true gap magnitude is at least γ_{ij}. The selected query is counted as diagnostic if its true gap magnitude is at least u_{ij}. Define the indicators below, where $\mathbf{1}\{E\}$ equals one if event E occurs and zero otherwise:

$$H_t := \mathbf{1}\{\mathcal{C}_t \cap \mathcal{X}_{I_t J_t}(\gamma_{I_t J_t}) \neq \varnothing\}, \qquad Z_t := \mathbf{1}\{x_t \in \mathcal{X}_{I_t J_t}(u_{I_t J_t})\},$$

$$q_t := \mathbb{E}[H_t \mid \mathcal{F}_{t-1}], \qquad Q_B := \sum_{t=1}^B q_t, \qquad L_B := \sum_{t=1}^B H_t(1 - Z_t), \qquad N_{\text{diag}}(B) := \sum_{t=1}^B Z_t.$$

Here q_t is the probability of a diagnostic opportunity conditional on the available history, and Q_B sums these probabilities over active rounds. The count L_B records missed opportunities, for which

$H_t = 1$ but $Z_t = 0$; $N_{\mathrm{diag}}(B)$ counts all diagnostic selected queries. These are analytical quantities because the true diagnostic regions are latent.

Theorem 1 (Diagnostic acquisition with missed opportunities)**.** *Fix an active budget $B \geq 1$ and diagnostic thresholds $0 < u_{ij} \leq \gamma_{ij}$. Suppose (I_t, J_t) is $\mathcal{F}_{t-1}$-measurable and each round's candidate set and selected query are measurable with respect to $\mathcal{F}_t$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$N_{\mathrm{diag}}(B) \geq Q_B - L_B - \sqrt{\frac{B}{2} \log \frac{1}{\delta}}.$$

The bound subtracts missed opportunities from cumulative conditional coverage; the square-root term accounts for fluctuations in candidate-set hits. For constants $\bar{p}, \varepsilon \in [0, 1]$ and $\eta \in (0, 1)$, if $Q_B \geq \bar{p}B$ and $L_B \leq \varepsilon B$ jointly with probability at least $1 - \eta$, then $N_{\mathrm{diag}}(B) \geq (\bar{p} - \varepsilon)B - \sqrt{(B/2)\log(1/\delta)}$ with probability at least $1 - \delta - \eta$. Here $\bar{p}$ lower-bounds average conditional coverage and ε upper-bounds the missed-opportunity count divided by B. Appendix A proves Theorem 1 and the bound above. For the Gaussian generator used in our simulation study, we provide sufficient conditions controlling both Q_B and L_B, yielding an explicit lower bound on the number of diagnostic queries selected for a fixed pair (Corollary 12).

For comparison, suppose DASE targets the same pair (i, j) in all B rounds. Random sampling from $\mathcal{D}$ encounters a diagnostic query with probability $\rho_{ij}(u_{ij})$ per draw. Suppose $Q_B \geq \bar{p}B$ and $L_B \leq \varepsilon B$ jointly with probability at least $1 - \eta$. If $\bar{p} - \varepsilon > \rho_{ij}(u_{ij})$, then DASE selects more diagnostic queries than B independent deployment draws with probability at least $1 - \delta - \eta$, provided B is large enough for the difference to exceed random fluctuations. This condition is easier to satisfy when diagnostic sparsity holds at threshold u_{ij} or candidate coverage $\bar{p}$ is high. Corollary 4 in Appendix A.3 makes this comparison precise.

4.2 Probability of correct selection

To connect diagnostic acquisition to the final decision, suppose the estimator admits an error bound that is nonincreasing in the number of diagnostic queries. More precisely, let $e_k := \mathbb{E}_{X \sim \mathcal{D}}[|\hat{\mu}_k(X) - \mu_k(X)|]$ be agent k's finite prediction error averaged over deployment queries, and write $M_B = N_{\mathrm{diag}}(B)$. Suppose $\max_k e_k \leq r_T(M_B, \eta)$ with probability at least $1 - \eta$, where $r_T(n, \eta)$ is a deterministic error bound that is nonincreasing in n and $\eta \in (0, 1)$ is its failure probability. Assumption 1 in Appendix B.1 formalizes this condition. If Theorem 1 provides a deterministic count lower bound n_0 with probability at least $1 - \alpha$, where $\alpha \in (0, 1)$, and $r_T(n_0, \eta) < \Delta_{\min}/2$, then $\mathbb{P}(\hat{k} = k^*) \geq 1 - \alpha - \eta$ (Theorem 6). Here $\Delta_{\min}$ is the minimum deployment-value gap defined in Section 2. Choosing $\alpha + \eta \leq \delta$ gives the target probability $1 - \delta$.

We make this connection explicit for NW estimation in a setting with finitely many possible queries. With warmup coverage of all query types, additional diagnostic queries reduce the estimation-error bound (Corollary 11). Combining this bound with diagnostic acquisition gives a correct-selection guarantee for a fixed agent pair (Theorem 13).

5 Experiments

We evaluate DASE in three settings: a synthetic sparse-signal problem, mathematical reasoning on the MATH benchmark, and clinical decision-making using hospital admissions.

Methods. We compare five methods. **Random** samples queries from $\mathcal{D}$. **TradQBC** (traditional query by committee; Seung et al., 1992) selects the query in a fixed pool $\mathcal{C}_{\mathrm{pool}}$ that maximizes disagreement among the agents' estimated conditional values, measured by $\mathrm{Var}_{k \in [K]}[\hat{\mu}_k(x)]$. **DASE Full** applies the DASE acquisition rule to the union of $\mathcal{C}_{\mathrm{pool}}$ and N_{gen} candidates generated at each round. **DASE Pool Only** uses the same rule on $\mathcal{C}_{\mathrm{pool}}$ alone, and **DASE Gen Only** uses it on the generated candidates alone. All methods use the NW estimator of Section 3 and share the same warmup queries. We estimate best-agent selection accuracy, $\mathbb{P}(\hat{k} = k^*)$, by the fraction of independent trials that return the best agent. Each budget T includes warmup and is evaluated in a separate run. Bold entries mark the highest accuracy for each pool and budget. Appendix D.1 gives the warmup budget, uncertainty-bonus coefficient, and other implementation settings. Generating new queries with an

Table 1: Best-agent selection accuracy in the synthetic study with Standard Pool (1,000 trials per budget) and the MATH study with Standard Pool (500 trials per budget). Tables 3 and 5 report all three pool configurations with standard errors.

	Synthetic, Standard				MATH, Standard			
Budget T	100	200	300	500	50	75	100	200
Random	0.559	0.624	0.715	0.801	0.592	0.674	0.722	0.838
TradQBC	0.498	0.565	0.587	0.643	0.602	0.656	0.686	0.760
DASE Pool Only	**0.596**	0.711	0.830	0.925	**0.660**	**0.756**	**0.826**	0.896
DASE Gen Only	0.568	**0.738**	**0.845**	**0.932**	0.646	0.702	0.760	0.844
DASE Full	0.567	0.709	0.805	0.930	0.628	0.754	0.810	**0.900**

LLM would require additional human validation and annotation, and the quality of the generated queries can be sensitive to the model's capabilities. In the MATH and clinical studies, DASE Full and DASE Gen Only therefore use retrieval from a separate pool as a proxy for LLM-based query generation. Using recorded agent outcomes also allows all methods to be evaluated on the same underlying data across repeated trials, supporting reproducible comparisons.

5.1 Synthetic sparse-signal study

Setup. We consider $K = 5$ agents on $\mathcal{X} = [0, 1]$, with a uniform deployment distribution and Bernoulli outcomes. Let $b_c(x) = \exp\big[-500(x - c)^2\big]$ denote a narrow bump centered at c. The best agent, $k^* = 3$, has conditional value $\mu_3(x) = 0.5 + 0.45\, b_{0.5}(x)$, whereas agents 1 and 2 have $\mu_k(x) = 0.5$. Agents 4 and 5 have lower deployment values, but their conditional values vary sharply near two other locations:

$$\mu_4(x) = 0.35 + 0.60\, b_{0.15}(x) - 0.30\, b_{0.85}(x), \qquad \mu_5(x) = 0.35 - 0.30\, b_{0.15}(x) + 0.60\, b_{0.85}(x).$$

The deployment values are $(V_1, \ldots, V_5) \approx (0.500, 0.500, 0.536, 0.374, 0.374)$, so $\Delta_{\min} \approx 0.036$. Selecting the best agent therefore requires detecting agent 3's advantage near $x = 0.5$. The largest differences among agents instead occur near $x = 0.15$ and $x = 0.85$ and involve agents with lower deployment values. A query sampled uniformly from $[0, 1]$ falls in the central interval $[0.47, 0.53]$ with probability 0.06.

Implementation. All methods use bandwidth $h = 0.04$, and Random samples uniformly from $[0, 1]$. DASE Gen Only and DASE Full generate $N_{\text{gen}} = 20$ candidates per round. Fifteen are Gaussian perturbations (standard deviation h) of evaluated queries with the largest estimated leader–challenger gaps. Five are sampled uniformly from $[0, 1]$. The fixed pool, **Standard Pool**, is a 50-point uniform grid. We evaluate $T \in \{100, 200, 300, 500\}$ using 1,000 trials per budget. Appendix D.2 reports two additional pools: **Small Pool**, a 10-point uniform grid, and **Restricted Pool**, which excludes the grid points in $[0.47, 0.53]$.

Results. The synthetic columns of Table 1 report selection accuracy. At $T = 500$, the three DASE variants achieve accuracies of 0.925 to 0.932, compared with 0.801 for Random and 0.643 for TradQBC. When the pool excludes the central interval (Restricted Pool), DASE Pool Only falls to 0.710 at the same budget, whereas DASE Gen Only and DASE Full reach 0.932 and 0.909, respectively (Table 3), showing the benefit of generation when the pool omits diagnostic queries.

5.2 MATH benchmark

Setup. We use the MATH competition dataset, which contains 12,500 problems (Hendrycks et al., 2021). The 5,000-problem test split represents the deployment distribution $\mathcal{D}$ and is never queried. We divide the 7,500-problem training split into a 2,500-problem query subset, from which warmup and Random's queries are sampled, and a disjoint 5,000-problem generation pool. Each problem is represented by an 11-dimensional feature vector for estimation and retrieval (Appendix D.1). The vector includes the problem's subject and reference-solution attributes, which are known before evaluation but do not directly reveal agent performance. The estimator uses bandwidth $h = 0.5$.

Agents. Four agents are prompting strategies applied to GPT-4o-mini: Direct, Zero-shot chain-of-thought (CoT) (Kojima et al., 2022), Four-shot CoT (Wei et al., 2022), and Self-Consistency (Wang et al., 2023) (Appendix E). To make selection more difficult, we construct a fifth agent that reduces the minimum deployment-value gap $\Delta_{\min}$. This agent, **NT-Diagnostic**, has recorded outcomes that match Self-Consistency outside Number Theory and Zero-shot CoT on Number Theory problems. The deployment values are $(V_1, \ldots, V_5) \approx (0.655, 0.655, 0.658, 0.718, 0.709)$, so $k^* = 4$ and $\Delta_{\min} \approx 0.0094$. Agents 4 and 5 differ only on Number Theory problems, where their mean outcomes differ by approximately 0.087.

Implementation. The generator retrieves existing problems using up to five evaluated problems with the largest estimated leader–challenger gaps as anchors. It forms $N_{\text{gen}} = 50$ candidates from the anchors' nearest neighbors and uniformly sampled problems in the generation pool. The fixed pool, **Standard Pool**, contains 500 problems sampled from the query subset with stratification by subject and difficulty. We evaluate $T \in \{50, 75, 100, 200\}$ using 500 trials per budget. Appendix D.3 reports two additional pools: **Small Pool**, a 100-problem pool, and **Restricted Pool**, which contains 500 problems sampled after excluding Number Theory.

Results. The MATH columns of Table 1 report selection accuracy. With Standard Pool, DASE Pool Only achieves higher accuracy than both baselines at every tested budget. At $T = 100$, it reaches 0.826, compared with 0.686 for TradQBC and 0.722 for Random. At $T = 200$, DASE Full and DASE Pool Only reach 0.900 and 0.896, compared with 0.760 for TradQBC and 0.838 for Random.

Under Restricted Pool at $T = 100$ (Table 5), TradQBC and DASE Pool Only reach 0.622 and 0.624, below Random at 0.722. Random samples from the full query subset and can therefore reach Number Theory problems. DASE Gen Only and DASE Full reach 0.760 and 0.774. As in the synthetic study, access to candidates outside the fixed pool improves selection when the pool excludes the region that distinguishes the leading agents.

5.3 Clinical decision-making

Setup. We use MIMIC-IV-Ext Clinical Decision Making (Hager et al., 2024), derived from MIMIC-IV (Johnson et al., 2023). The dataset contains hospital admissions for acute abdominal pain labeled as appendicitis, cholecystitis, pancreatitis, or diverticulitis. Each agent evaluation covers the full interaction, from the initial history through information requests to a final diagnosis, scored against the reference diagnosis (Figure 2A; Appendix F.1). We use 899 admissions with precomputed agent rollouts. A fixed reference set D of 500 admissions defines the deployment values V_k and best agent k^* and is never queried. Each trial divides the remaining 399 admissions into a query subset of 200 and a disjoint generation pool of 199, holding D fixed (Appendix F.4). Warmup, Random, and the fixed pools draw from the query subset. The nine-dimensional feature vector for estimation and retrieval includes the known pathology label (Appendix F.5).

Agents. We evaluate $K = 6$ agents by pairing Llama-3.1-8B-Instruct (Dubey et al., 2024) and the medically tuned OpenBioLLM-70B (Pal & Sankarasubbu, 2024) with three protocols for requesting information. P0 is the unmodified baseline, P1 requires a ranked differential diagnosis and a justification for each request, and P2 limits the agent to five requests (Appendix F.2). We serve both models locally on institution-owned hardware, following PhysioNet's recommendation (PhysioNet, 2025). Figure 2B shows reference-set accuracies. The best agent is Llama-3.1-8B with P2, with $V_{k^*} \approx 0.6372$, followed by P0 at 0.6196, a gap of 0.0176. Most of this gap comes from cholecystitis cases.

Implementation. The estimator uses bandwidth $h = 0.5$. The generator retrieves admissions with recorded outcomes from a separate pool, using evaluated admissions with large estimated leader–challenger gaps as anchors (Appendix F.6). We evaluate three fixed pools: **Standard Pool**, the entire query subset; **Small Pool**, a stratified subsample of 25 admissions per pathology; and **Restricted Pool**, the 150 query-subset admissions remaining after excluding cholecystitis. We evaluate $T \in \{90, 120, 150\}$ using 1,000 trials per budget. To account for the generator's access to additional queries, we add a baseline that samples uniformly from the union of the query subset and the generation pool (Appendix D.4). We also evaluate LLM-guided DASE variants at $T = 120$ over 200 trials

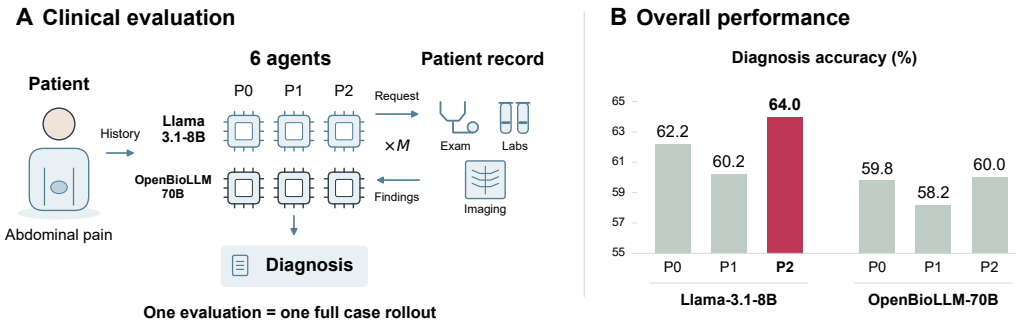

Figure 2: **Clinical evaluation and agent performance.** **(A)** Each evaluation is a full case rollout with M information requests. Six agents combine two models with three protocols. **(B)** Recorded diagnosis accuracy on 500 reference admissions.

Table 2: Best-agent selection accuracy in the clinical study from 1,000 trials per pool and budget. The query subset, generation pool, and fixed pools are redrawn in each trial. Random and DASE Gen Only do not use the fixed pool.

	Standard			Small			Restricted		
Budget T	90	120	150	90	120	150	90	120	150
Random	0.327	0.340	0.369	0.327	0.340	0.369	0.327	0.340	0.369
TradQBC	0.333	0.349	0.381	0.329	0.353	0.359	0.242	0.285	0.295
DASE Pool Only	0.380	0.433	0.436	0.328	0.349	0.348	0.288	0.293	0.315
DASE Gen Only	0.362	0.398	0.421	0.362	0.398	0.421	**0.362**	**0.398**	**0.421**
DASE Full	**0.391**	**0.458**	**0.484**	**0.383**	**0.429**	**0.451**	0.335	0.375	0.409

per pool, using a locally served Qwen2.5-32B model to select admissions from retrieved candidates (Appendix F.7).

Results. Table 2 reports selection accuracy. DASE Full has the highest accuracy under Standard Pool and Small Pool at each budget reported in the table. At $T = 150$, it reaches 0.484 under Standard Pool, compared with 0.436 for DASE Pool Only, 0.381 for TradQBC, and 0.369 for Random. Under Small Pool, DASE Full reaches 0.451, compared with 0.348 for DASE Pool Only and 0.359 for TradQBC, both of which exhaust the available pool queries before reaching the budget. At this budget, DASE Full allocates 47% of its active queries to generated candidates under Standard Pool and 65% under Small Pool, relying more on generation when the fixed pool is smaller.

Under Restricted Pool, TradQBC and DASE Pool Only have lower accuracy than Random at every reported budget. At $T = 150$, they reach 0.295 and 0.315, compared with 0.369 for Random. This pool excludes cholecystitis cases, which account for most of the gap between the leading agents. Generation restores access to these admissions: DASE Gen Only reaches 0.421 and DASE Full reaches 0.409. DASE also supports LLM-guided query proposal: with a local model selecting from retrieved candidates, DASE Full (LLM) achieves higher selection accuracy than TradQBC and DASE Pool Only under all three pools (Appendix F.7).

6 Conclusion

We introduced DASE, which selects among fixed agents under a limited budget by generating and selecting queries to distinguish the current leader and challenger. We analyze how many diagnostic queries DASE selects and when it can identify the best agent. Across three studies, DASE generally improves selection accuracy over the tested baselines, with generation helping when fixed pools omit queries that distinguish leading agents. Future work will extend DASE to choose both the query and the subset of agents evaluated at each round. This could reduce evaluation costs by focusing on competitive agents while retaining enough evidence for reliable selection.

AI USE STATEMENT

We used AI tools to assist with writing and editing the manuscript, code implementation, and theoretical derivations. The authors take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide code for DASE and the comparison methods, together with instructions for reproducing the synthetic experiments, in the supplementary material. Implementation details for the MATH and clinical studies are documented in Appendices E and F.

REFERENCES

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002. doi: 10.1023/A:1013689704352.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yavuz Durmazkeser, Patrik Okanovic, Andreas Kirsch, Torsten Hoefler, and Nezihe Merve Gürel. Active model selection for large language models. *arXiv preprint arXiv:2510.09418*, 2025.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

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 A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 2016.

Douwe Kiela, Max Bartolo, Yixin Nie, et al. Dynabench: Rethinking benchmarking in NLP. *North American Chapter of the Association for Computational Linguistics*, 2021.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5753–5763, 2021.

Xiang Lisa Li, Farzaan Kaiyom, Evan Zheran Liu, Yifan Mai, Percy Liang, and Tatsunori Hashimoto. AutoBencher: Towards declarative benchmark construction. In *International Conference on Learning Representations*, 2025.

Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. SymPy: symbolic computing in Python. *PeerJ Computer Science*, 3:e103, 2017.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964. doi: 10.1137/1109020.

Ankit Pal and Malaikannan Sankarasubbu. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B`, 2024.

PhysioNet. Use of MIMIC data with large language models and online services. `https://physionet.org/news/post/llm-responsible-use/`, 2025.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Daniel Russo. Simple Bayesian algorithms for best arm identification. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49, pp. 1417–1418, 2016.

Christoph Sawade, Niels Landwehr, and Tobias Scheffer. Active comparison of prediction models. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. *Conference on Learning Theory*, 1992.

Elad Tolochinsky, Yaniv Tenzer, and Yaniv Romano. Valid best-model identification for LLM evaluation via low-rank factorization. *arXiv preprint arXiv:2605.10405*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26: 359–372, 1964. URL `https://www.jstor.org/stable/25049340`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

A Proofs for Diagnostic Acquisition

Proof of Theorem 1. By definition, $H_t - q_t$ is a martingale difference with a conditional range of length one. Conditional Hoeffding's lemma and the exponential form of Markov's inequality give, for $a > 0$,

$$\mathbb{P}\left(\sum_{t=1}^{B}(H_t - q_t) \le -a\right) \le \inf_{\lambda>0} \exp(-\lambda a + B\lambda^2/8) = \exp(-2a^2/B).$$

For every realization, $Z_t \ge H_t - H_t(1 - Z_t)$, so

$$N_{\mathrm{diag}}(B) \ge \sum_{t=1}^{B} H_t - L_B.$$

Combining the two inequalities with $a = \sqrt{(B/2)\log(1/\delta)}$ proves the theorem. Intersecting this event with $\{Q_B \ge \bar{p}B,\ L_B \le \varepsilon B\}$ and applying a union bound gives the cumulative-rate bound stated after the theorem, without requiring independence between the events. □

A.1 Controlling missed opportunities

Write the acquisition score in equation 1 with explicit round indices as

$$A_t(x) := |\hat{\mu}_{I_t,t}(x) - \hat{\mu}_{J_t,t}(x)| + \beta_t(\hat{\sigma}_{I_t,t}(x) + \hat{\sigma}_{J_t,t}(x)).$$

Let $d_t(x) := |D_{I_tJ_t}(x)|$ be the true gap magnitude for the target pair. In a round with $H_t = 1$, suppose a candidate x^* satisfies $d_t(x^*) \ge \gamma_{I_tJ_t}$ and the following score bounds hold:

$$A_t(x^*) \ge d_t(x^*) - a_t, \qquad A_t(x_t) \le d_t(x_t) + b_t, \qquad a_t, b_t \ge 0.$$

Here a_t bounds score underestimation at x^*, and b_t bounds score overestimation at the selected query x_t. Maximizing A_t gives

$$d_t(x_t) + b_t \geq A_t(x_t) \geq A_t(x^*) \geq d_t(x^*) - a_t \geq \gamma_{I_t J_t} - a_t.$$

Hence this round contributes nothing to L_B whenever $a_t + b_t \leq \gamma_{I_t J_t} - u_{I_t J_t}$. If the criterion fails on at most m rounds with $H_t = 1$, then $L_B \leq m$.

For NW estimation with bounded outcomes, the same criterion applies to the empirical uncertainty proxy in Section 3 without assuming that it is a calibrated confidence radius. Let $W_t^{\mathrm{pre}}(x)$ denote the total kernel mass from observations available before active round t, including warmup.

Lemma 2 (A sufficient condition for NW acquisition)**.** *Suppose outcomes lie in* $[0,1]$*, all agents use the same NW weights, and the kernel takes values in* $[0,1]$*. Fix an active round* t*. Suppose that, at every candidate* x*, each target agent's conditional-value estimate has absolute error at most* $\epsilon_t \geq 0$ *and* $W_t^{\mathrm{pre}}(x) \geq w_t > 0$*. If*

$$4\epsilon_t + \frac{\beta_t}{\sigma\sqrt{w_t}} \leq \gamma_{I_t J_t} - u_{I_t J_t}, \tag{2}$$

then $H_t = 1$ *implies* $Z_t = 1$*. Here* σ *is the scale parameter in the empirical uncertainty proxy.*

Proof of Lemma 2. Let $A_{t,s}^{\mathrm{pre}}(x)$ denote the normalized NW weight of observation s before active round t, with sums taken over all observations then available. The weighted variance of outcomes in $[0,1]$ is at most $1/4$. Since the kernel is bounded by one, $\sum_s [A_{t,s}^{\mathrm{pre}}(x)]^2 \leq 1/W_t^{\mathrm{pre}}(x)$. The empirical uncertainty proxy therefore satisfies

$$\hat{\sigma}_{I_t,t}(x) + \hat{\sigma}_{J_t,t}(x) \leq \frac{1}{\sigma\sqrt{W_t^{\mathrm{pre}}(x)}} \leq \frac{1}{\sigma\sqrt{w_t}}.$$

The estimated absolute pair gap differs from its true magnitude by at most $2\epsilon_t$. Nonnegativity of the bonus gives the local score bounds with $a_t = 2\epsilon_t$ and $b_t = 2\epsilon_t + \beta_t/(\sigma\sqrt{w_t})$. The local acquisition criterion and equation 2 prove the claim. □

Consequently, $L_B = 0$ when these conditions hold in every active round. If they fail on at most m rounds with $H_t = 1$, then $L_B \leq m$. Appendices C.2 and C.3 give simultaneous estimation bounds that can be used in this criterion.

A.2 Candidate coverage

If the N generated candidates are conditionally independent and each has probability at least $\alpha_t \in [0,1]$ of entering $\mathcal{X}_{I_t J_t}(\gamma_{I_t J_t})$, conditional on $\mathcal{F}_{t-1}$, then $q_t \geq 1 - (1-\alpha_t)^N$. Without conditional independence, Theorem 1 still applies using q_t. For a given history and target pair, adding a fixed pool can increase candidate coverage and change the selected query. The theorem applies to both DASE Gen Only and DASE Full through their respective Q_B and L_B.

For the Gaussian generator in Section 5.1, we can bound conditional coverage explicitly. Fix a target pair (i,j) throughout the active phase, write $d(x) = |D_{ij}(x)|$, and let $\mathcal{D}$ be uniform on $[0,1]$. For brevity, write $\gamma = \gamma_{ij}$ and $u = u_{ij}$. Suppose $J = [a,b] \subseteq \mathcal{X}_{ij}(\gamma)$, and write $|J| = b - a$ for its length. Conditional on $\mathcal{F}_{t-1}$, draw N_0 uniform candidates and r_ℓ candidates from a Gaussian centered at anchor z_ℓ, with standard deviation h_g, truncated to $[0,1]$. The anchors and allocations are determined by $\mathcal{F}_{t-1}$, and all candidate draws are conditionally independent. Let Φ denote the standard normal cumulative distribution function. A Gaussian draw centered at z lies in J with probability

$$p_J(z) = \frac{\Phi((b-z)/h_g) - \Phi((a-z)/h_g)}{\Phi((1-z)/h_g) - \Phi(-z/h_g)}. \tag{3}$$

Independence therefore gives

$$\mathbb{P}(\mathcal{C}_{\mathrm{gen},t} \cap J \neq \varnothing \mid \mathcal{F}_{t-1}) = 1 - (1-|J|)^{N_0} \prod_\ell (1 - p_J(z_\ell))^{r_\ell}.$$

This probability lower-bounds q_t, including when a fixed pool is added. Choose $0 < d_\circ < (b-a)/2$ and define the interior interval $J_\circ = [a + d_\circ, b - d_\circ]$. If $z \in J_\circ$, the interval $[z - d_\circ, z + d_\circ]$ lies in J,

so equation 3 implies $p_J(z) \geq c_\circ := 2\Phi(d_\circ/h_g) - 1$. If at least one anchor lies in $J_\circ$ and receives at least r draws, then

$$q_t \geq p_* := 1 - (1 - |J|)^{N_0}(1 - c_\circ)^r. \tag{4}$$

The synthetic generator uses $N_0 = 5$ and divides fifteen local draws among at most five anchors, giving $r \geq 3$ for each selected anchor after warmup.

Proposition 3 (Retention of a diagnostic anchor)**.** *Suppose $S \subset J_\circ$ has uniform probability $s > 0$ and there are constants $v, \zeta > 0$ such that*

$$\inf_{x \in S} d(x) \geq v, \qquad \sup_{x \notin J_\circ} d(x) \leq v - \zeta. \tag{5}$$

Warmup consists of T_0 independent uniform queries. All evaluated queries remain in history, and the anchor set includes a maximizer of the estimated absolute pair gap over history. On any event where both agents' conditional-value estimates have absolute error at most $\epsilon \geq 0$ at every evaluated query in every active round, $4\epsilon < \zeta$ ensures that a warmup query in S yields an anchor in $J_\circ$ in every round. The probability that warmup misses S is $(1 - s)^{T_0}$.

Proof of Proposition 3. A warmup query in S has estimated absolute gap at least $v - 2\epsilon$. Each point outside $J_\circ$ has estimated gap at most $v - \zeta + 2\epsilon < v - 2\epsilon$. A maximizer over history must therefore lie in $J_\circ$. The warmup query remains in history, so the argument applies in every round. Independent uniform warmup gives the stated probability of missing S. □

For either synthetic pair $(3, 1)$ or $(3, 2)$, $d(x) = 0.45e^{-500(x-1/2)^2}$. Choose $0 < s_0 < c - d_\circ < c < 1/2$, $J = [1/2 - c, 1/2 + c]$, and $S = [1/2 - s_0, 1/2 + s_0]$. The conditions above hold with

$$\gamma = 0.45e^{-500c^2}, \quad s = 2s_0, \quad v = 0.45e^{-500s_0^2}, \quad \zeta = v - 0.45e^{-500(c-d_\circ)^2} > 0.$$

These choices satisfy equation 5 and make equation 4 explicit for a fixed pair. Appendix C.3 controls the required estimation error and combines this coverage bound with Lemma 2 and Theorem 1.

A.3 Comparison with random sampling

Corollary 4 (Fixed-pair diagnostic-count comparison)**.** *Fix (i, j) and suppose DASE targets this pair in all B rounds. Let $\rho_{ij}^u := \rho_{ij}(u_{ij})$ be the deployment probability of a diagnostic query at threshold u_{ij}, and write $N_{\text{diag}}^{\text{DASE}}(B) := N_{\text{diag}}(B)$. Let $N_{\text{diag}}^{\text{Random}}(B)$ count queries in $\mathcal{X}_{ij}(u_{ij})$ among B independent draws from $\mathcal{D}$. For any $\delta \in (0, 1)$, with joint probability at least $1 - \delta$,*

$$N_{\text{diag}}^{\text{DASE}}(B) - N_{\text{diag}}^{\text{Random}}(B) \geq Q_B - L_B - \rho_{ij}^u B - 2\sqrt{\frac{B}{2}\log\frac{2}{\delta}}.$$

Proof of Corollary 4. The indicators of diagnostic queries under Random are independent Bernoulli variables with mean ρ_{ij}^u. Hoeffding's inequality gives, with probability at least $1 - \delta/2$,

$$N_{\text{diag}}^{\text{Random}}(B) \leq \rho_{ij}^u B + \sqrt{\frac{B}{2}\log\frac{2}{\delta}}.$$

Apply Theorem 1 at failure probability $\delta/2$ and intersect the two events. Subtraction proves the claim; the procedures need not be independent of each other. □

B Selection Regret from Integrated Estimation Error

Proposition 5 (Selection regret from integrated estimation error)**.** *For final conditional-value estimates $\hat{\mu}_k$, let $\hat{V}_k = \mathbb{E}_{X \sim \mathcal{D}}[\hat{\mu}_k(X)]$ and $\hat{k} \in \arg\max_k \hat{V}_k$. Define the deployment-integrated absolute errors and assume they are finite:*

$$e_k := \mathbb{E}_{X \sim \mathcal{D}}[|\hat{\mu}_k(X) - \mu_k(X)|].$$

Then, for every realized history,

$$V_{k^*} - V_{\hat{k}} \leq e_{k^*} + e_{\hat{k}} \leq 2\max_k e_k.$$

In particular, $\max_k e_k < \Delta_{\min}/2$ *implies* $\hat{k} = k^*$. *If* $\mathbb{E}[\max_k e_k] \le \epsilon$ *for a constant* $\epsilon \ge 0$*, where the outer expectation is over evaluation histories and algorithmic randomness, then*

$$\mathbb{E}[V_{k^*} - V_{\hat{k}}] \le 2\epsilon, \qquad \mathbb{P}(\hat{k} \ne k^*) \le \min\{1, 2\epsilon/\Delta_{\min}\}.$$

Proof of Proposition 5. Integration and the triangle inequality give $|\hat{V}_k - V_k| \le e_k$. Since $\hat{V}_{\hat{k}} \ge \hat{V}_{k^*}$,

$$V_{k^*} - V_{\hat{k}} = (V_{k^*} - \hat{V}_{k^*}) + (\hat{V}_{k^*} - \hat{V}_{\hat{k}}) + (\hat{V}_{\hat{k}} - V_{\hat{k}}) \le e_{k^*} + e_{\hat{k}} \le 2 \max_k e_k.$$

Every incorrect selection incurs regret at least $\Delta_{\min}$. This establishes correct selection when the maximum integrated error is strictly below half the minimum value gap. Taking expectations gives the expected-regret bound, and

$$\Delta_{\min}\mathbb{P}(\hat{k} \ne k^*) \le \mathbb{E}[V_{k^*} - V_{\hat{k}}] \le 2\epsilon$$

proves the probability bound. □

Remark 1 (Diagnostic counts and deployment values)**.** *Suppose* $D_{12}(x) = 1$ *on a region of deployment mass* 0.1 *and* $D_{12}(x) = -0.2$ *elsewhere. Then* $V_1 - V_2 = -0.08$*, while only the first region is diagnostic at threshold* 0.5 *and favors agent 1. Thus diagnostic counts alone do not ensure correct selection; estimation error outside the diagnostic region must also be controlled.*

B.1 An estimator condition linking acquisition to selection

Let $M_B := N_{\mathrm{diag}}(B)$ count the selected diagnostic queries. The following condition depends jointly on the estimator, query-selection procedure, and data distribution.

Assumption 1 (Integrated error controlled by diagnostic count)**.** *For the fixed budget* $T = T_0 + B$*, there is a deterministic function* $r_T : [0, B] \times (0, 1) \to [0, \infty]$*, nonincreasing in its first argument, such that, for every* $\eta \in (0, 1)$*,*

$$\mathbb{P}\left(\max_k e_k \le r_T(M_B, \eta)\right) \ge 1 - \eta.$$

The errors e_k *are defined in Proposition 5. Probability is over evaluation histories and algorithmic randomness. The value* $r_T(0, \eta) = \infty$ *is allowed.*

The probability statement is evaluated at the random count M_B and does not condition on $M_B = n$. It does not require the empirical uncertainty proxy $\hat{\sigma}_k$ to be a calibrated confidence radius.

Theorem 6 (Correct selection from a diagnostic-count bound)**.** *Suppose Assumption 1 holds. Let* $n_0 \in [0, B]$ *be deterministic and suppose* $\mathbb{P}(M_B \ge n_0) \ge 1 - \alpha$*, where* $\alpha \in (0, 1)$*. For any* $\eta \in (0, 1)$*, if* $r_T(n_0, \eta) < \Delta_{\min}/2$*, then*

$$\mathbb{P}(\hat{k} = k^*) \ge 1 - \alpha - \eta.$$

Proof. With probability at least $1 - \alpha - \eta$, both $M_B \ge n_0$ and $\max_k e_k \le r_T(M_B, \eta)$ hold. Monotonicity then gives $\max_k e_k \le r_T(n_0, \eta) < \Delta_{\min}/2$, which implies correct selection by Proposition 5. The two events need not be independent. □

To apply Theorem 1, suppose $Q_B \ge \bar{p}B$ and $L_B \le \varepsilon B$ jointly with probability at least $1 - \zeta$, for deterministic $\bar{p}, \varepsilon \in [0, 1]$ and $\zeta \in (0, 1)$. For $\xi \in (0, 1)$, define

$$n_0 := \max\left\{0, (\bar{p} - \varepsilon)B - \sqrt{\frac{B}{2}\log\frac{1}{\xi}}\right\}.$$

The acquisition theorem gives $\mathbb{P}(M_B \ge n_0) \ge 1 - \xi - \zeta$. Thus $r_T(n_0, \eta) < \Delta_{\min}/2$ ensures $\mathbb{P}(\hat{k} = k^*) \ge 1 - \xi - \zeta - \eta$. Appendix C.1 verifies Assumption 1 for NW under explicit conditions on deployment-integrated kernel coverage and smoothing bias.

C NW Error under Adaptive Query Locations

We first bound the integrated error of the NW estimator using realized kernel coverage and smoothing bias under adaptive query selection. We then verify the estimator condition in Assumption 1 and derive explicit bounds for finite query spaces and Gaussian generation.

Assumption 2 (Noise and kernel for the NW error bound)**.** *Index all evaluations, including warmup, by $s = 1, \dots, T$, for a fixed total budget $T \geq 1$. Let $\mathcal{P}_s$ contain all information available immediately before the outcomes at x_s are observed, including x_s; it includes all earlier outcomes. For each k and $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\big[e^{\lambda \xi_{k,s}} \mid \mathcal{P}_s\big] \leq e^{\lambda^2 \nu^2/2}, \qquad \xi_{k,s} := Y_k(x_s) - \mu_k(x_s), \quad \nu > 0.$$

The measurable kernel $K_h : \mathcal{X} \times \mathcal{X} \to [0,1]$, including its bandwidth, is fixed before evaluation. The deployment distribution is fixed independently of the evaluation noise, and the conditional means are deployment-integrable. All quantities below are measurable. The kernel mass $W_T(x)$ defined below is positive for $\mathcal{D}$-almost every x, almost surely.

For the final NW fit, define

$$W_T(x) := \sum_{s=1}^{T} K_h(x, x_s), \qquad U_T(x) := \sum_{s=1}^{T} K_h(x, x_s)^2, \qquad A_{T,s}(x) := \frac{K_h(x, x_s)}{W_T(x)},$$

$$\hat{\mu}_k(x) := \sum_{s=1}^{T} A_{T,s}(x) Y_k(x_s), \qquad \mathcal{B}_{k,T} := \int \sum_{s=1}^{T} A_{T,s}(x) |\mu_k(x_s) - \mu_k(x)| \, d\mathcal{D}(x),$$

$$\mathcal{C}_T := \int \frac{1 + U_T(x)}{W_T(x)^2} \, d\mathcal{D}(x), \qquad \ell_T(\eta) := \log \frac{K\sqrt{1+T}}{\eta}.$$

The quantity $\mathcal{B}_{k,T}$ bounds smoothing bias, and $\mathcal{C}_T$ measures inverse kernel coverage averaged over the deployment distribution. The added 1 is used only in the analysis and does not change the NW estimator.

Proposition 7 (Deployment-integrated NW error)**.** *Under Assumption 2, for every $\eta \in (0,1)$, with probability at least $1 - \eta$, simultaneously for all k,*

$$e_k \leq \mathcal{B}_{k,T} + \nu \sqrt{2 \ell_T(\eta) \mathcal{C}_T}.$$

The bound is informative when its right-hand side is finite. Define

$$R_T(\eta) := \max_k \mathcal{B}_{k,T} + \nu \sqrt{2 \ell_T(\eta) \mathcal{C}_T}.$$

On the same event, selecting the agent with the highest integrated NW estimate gives $V_{k^*} - V_{\hat{k}} \leq 2R_T(\eta)$. In particular,

$$\mathbb{P}\big(\hat{k} \neq k^*, \; 2R_T(\eta) < \Delta_{\min}\big) \leq \eta.$$

Proof of Proposition 7. For fixed x and k, let $w_s(x) := K_h(x, x_s)$ and $S_{k,T}(x) := \sum_{s=1}^{T} w_s(x) \xi_{k,s}$. Each unnormalized weight $w_s(x)$ is known before the corresponding outcome is observed. Iterated conditional expectation under Assumption 2 therefore gives, for every $\theta \in \mathbb{R}$,

$$\mathbb{E} \exp\big(\theta S_{k,T}(x)/\nu - \theta^2 U_T(x)/2\big) \leq 1.$$

Integrate this inequality against the standard normal density in θ. Tonelli's theorem and completion of the square yield

$$M_{k,T}(x) := \frac{1}{\sqrt{1 + U_T(x)}} \exp\left(\frac{S_{k,T}(x)^2}{2\nu^2(1 + U_T(x))}\right), \qquad \mathbb{E}[M_{k,T}(x)] \leq 1.$$

This is the scalar Gaussian-mixture construction used in Abbasi-Yadkori, Pál, and Szepesvári (2011); the remaining steps control a deployment integral rather than a pointwise confidence radius.

Since $\mathcal{D}$ is fixed, Tonelli gives $\mathbb{E}[\int M_{k,T}(x)\, d\mathcal{D}(x)] \le 1$. Markov's inequality and a union bound over k give an event of probability at least $1-\eta$ on which

$$\int M_{k,T}(x)\, d\mathcal{D}(x) \le K/\eta \quad \text{for all } k.$$

On this event, use $U_T(x) \le T$ and set $z_k(x) := S_{k,T}(x)^2/[2\nu^2(1+U_T(x))]$. Then

$$\int e^{z_k(x)}\, d\mathcal{D}(x) \le \sqrt{1+T} \int M_{k,T}(x)\, d\mathcal{D}(x) \le \frac{K\sqrt{1+T}}{\eta}.$$

Jensen's inequality, with truncation if needed, implies $\int z_k(x)\, d\mathcal{D}(x) \le \ell_T(\eta)$. Consequently, Cauchy–Schwarz gives

$$\begin{aligned}\int \frac{|S_{k,T}(x)|}{W_T(x)}\, d\mathcal{D}(x) &= \nu\sqrt{2} \int \sqrt{z_k(x)}\, \frac{\sqrt{1+U_T(x)}}{W_T(x)}\, d\mathcal{D}(x) \\ &\le \nu\sqrt{2\ell_T(\eta)\mathcal{C}_T}.\end{aligned}$$

Finally, the NW error decomposes pathwise as

$$\hat{\mu}_k(x) - \mu_k(x) = \sum_{s=1}^{T} A_{T,s}(x)(\mu_k(x_s) - \mu_k(x)) + \frac{S_{k,T}(x)}{W_T(x)}.$$

Integration and the triangle inequality give the error bound. If the right-hand side is infinite, the inequality holds in the extended sense but is uninformative. Otherwise, Proposition 5 gives the regret bound and ensures correct selection whenever $2R_T(\eta) < \Delta_{\min}$. The probability of incorrect selection together with this last inequality is therefore at most η. □

Assumption 2 allows adaptive query locations and independent or dependent agent outcomes, but fixes the kernel and bandwidth before evaluation. The bound controls deployment-integrated error rather than error uniformly over all query locations. For outcomes in $[0,1]$, one may take $\nu = 1/2$; this noise bound does not calibrate the empirical uncertainty proxy. Convergence rates or comparisons between sampling procedures require additional control of coverage and smoothing bias.

C.1 Verifying the estimator condition for NW

We verify Assumption 1 using two quantities averaged over the deployment distribution: the integrated NW smoothing bias and the second moment of inverse kernel mass, defined as

$$\widetilde{\mathcal{B}}_{k,T} := \int \left| \sum_{s=1}^{T} A_{T,s}(x)(\mu_k(x_s) - \mu_k(x)) \right| d\mathcal{D}(x), \qquad \mathcal{J}_T := \int W_T(x)^{-2}\, d\mathcal{D}(x).$$

The bias measures the error from averaging conditional means across queries. The quantity $\mathcal{J}_T$ is large when deployment queries receive little kernel mass. The triangle inequality gives $\widetilde{\mathcal{B}}_{k,T} \le \mathcal{B}_{k,T}$.

Proposition 8 (NW verification under integrated coverage and bias bounds)**.** *Suppose Assumption 2 holds and the final integrated errors are finite almost surely. For each $\eta \in (0,1)$, let $b_T(\cdot,\eta)$ and $v_T(\cdot,\eta)$ be deterministic, finite, nonnegative, nonincreasing functions on $(0,B]$. Suppose*

$$\mathbb{P}\left(M_B > 0,\ \left\{ \max_k \widetilde{\mathcal{B}}_{k,T} > b_T(M_B,\eta) \right\} \cup \{ \mathcal{J}_T > v_T(M_B,\eta) \} \right) \le \eta/2.$$

Then NW satisfies Assumption 1 with $r_T(0,\eta) = \infty$ and, for $n > 0$,

$$r_T(n,\eta) := b_T(n,\eta) + \nu\sqrt{2\ell_T(\eta/2)\left[\sqrt{v_T(n,\eta)} + v_T(n,\eta)\right]}, \qquad \ell_T(\eta/2) = \log \frac{2K\sqrt{1+T}}{\eta}.$$

Proof of Proposition 8. Since $0 \le K_h \le 1$, $U_T(x) \le W_T(x)$. Cauchy–Schwarz under the probability distribution $\mathcal{D}$ gives

$$\mathcal{C}_T \le \int \left[W_T(x)^{-1} + W_T(x)^{-2}\right] d\mathcal{D}(x) \le \sqrt{\mathcal{J}_T} + \mathcal{J}_T.$$

The noise argument in the proof of Proposition 7, applied with failure probability $\eta/2$, gives simultaneously for all k

$$\int \frac{|S_{k,T}(x)|}{W_T(x)}\, d\mathcal{D}(x) \le \nu\sqrt{2\ell_T(\eta/2)\mathcal{C}_T}, \qquad S_{k,T}(x) := \sum_{s=1}^{T} K_h(x, x_s)(Y_k(x_s) - \mu_k(x_s)).$$

The exact NW decomposition and the triangle inequality imply

$$e_k \le \widetilde{\mathcal{B}}_{k,T} + \int \frac{|S_{k,T}(x)|}{W_T(x)}\, d\mathcal{D}(x).$$

Intersect the noise event with the assumed event controlling integrated coverage and bias. A union bound limits the total failure probability to η, without requiring independence. On this intersection, $M_B > 0$ implies $\max_k e_k \le r_T(M_B, \eta)$. For $M_B = 0$, the claim follows from $r_T(0, \eta) = \infty$. The monotonicity of b_T and v_T implies that of r_T. □

The required bounds control deployment-integrated coverage and bias, without a common pointwise lower bound on kernel mass or global smoothness. They may fail with probability at most $\eta/2$ on histories with $M_B > 0$; the remaining $\eta/2$ is allocated to outcome noise. Correct selection further requires $r_T(n_0, \eta) < \Delta_{\min}/2$, as in Theorem 6.

C.2 An explicit error bound from diagnostic counts

On a finite query space, a simultaneous NW error bound gives an explicit function of the diagnostic count in Assumption 1. Throughout this subsection, Assumption 2 holds and outcomes lie in $[0, 1]$. We index all evaluations, including warmup, by s, and write $\hat{\mu}_{k,n}$ for the estimate after n evaluations, where $1 \le n \le T$. The quantities W_n, U_n, and $A_{n,s}$ are defined as in Proposition 7, with T replaced by n.

Lemma 9 (Simultaneous NW noise bound)**.** *Let $\mathcal{Z}$ be a deterministic set of $G \ge 1$ query locations with $W_n(z) > 0$ for every $z \in \mathcal{Z}$ and $1 \le n \le T$. For $\eta \in (0, 1)$, set $\Lambda_G(\eta) = \log(KGT\sqrt{1+T}/\eta)$. With probability at least $1 - \eta$, simultaneously for all k, n and $z \in \mathcal{Z}$,*

$$\frac{|S_{k,n}(z)|}{W_n(z)} \le \nu\frac{\sqrt{2\Lambda_G(\eta)(1 + U_n(z))}}{W_n(z)}, \qquad S_{k,n}(z) = \sum_{s=1}^{n} K_h(z, x_s)\xi_{k,s}. \tag{6}$$

Proof of Lemma 9. The Gaussian-mixture argument in the proof of Proposition 7 gives, for each fixed k, n, z,

$$\mathbb{E}\left[\frac{1}{\sqrt{1+U_n(z)}}\exp\left(\frac{S_{k,n}(z)^2}{2\nu^2(1+U_n(z))}\right)\right] \le 1.$$

Apply Markov's inequality at level KGT/η and take a union bound over k, n, z. Since $U_n(z) \le T$, rearranging yields equation 6. The query locations may be adaptive because each unnormalized kernel weight is known before its outcome is observed. □

To apply this bound on a finite query space, let $\mathcal{X} = \{z_1, \ldots, z_m\}$ consist of $m \ge 2$ distinct query types with deployment probabilities $p_j > 0$. Each evaluation selects one type, and repeated evaluations satisfy the conditional noise assumption. Use $K_h(z, z') = \exp(-\|z - z'\|^2/(2h^2))$ with fixed $h > 0$. Define

$$d_0 = \min_{j \ne \ell} \|z_j - z_\ell\| > 0, \qquad \kappa = e^{-d_0^2/(2h^2)}, \qquad n_j(n) = \sum_{s=1}^{n} \mathbf{1}\{x_s = z_j\}.$$

Fix positive integer warmup counts $a_1, \ldots, a_m$ before evaluation, with $\sum_j a_j = T_0$, and evaluate type j exactly a_j times during warmup. This allocation provides coverage of every deployment type. For a fixed $\eta \in (0, 1)$, abbreviate $\Lambda_m(\eta) = \log(KmT\sqrt{1+T}/\eta)$ as Λ_m.

Proposition 10 (NW error on a finite query space)**.** *With probability at least* $1-\eta$*, simultaneously for all* k, j *and* $T_0 \le n \le T$*,*

$$|\hat{\mu}_{k,n}(z_j) - \mu_k(z_j)| \le f_n(n_j(n)), \qquad f_n(v) = \kappa\left(\frac{n}{v} - 1\right) + \nu\sqrt{2\Lambda_m\left(\frac{1}{v} + \frac{1}{v^2}\right)}. \tag{7}$$

On the same event, $\max_k e_k \le \sum_j p_j f_T(n_j(T))$*.*

Proof of Proposition 10. At z_j, observations of type j have kernel weight one; all other weights are at most κ. Hence $W_n(z_j) \ge n_j(n)$, and the absolute smoothing bias is at most $\kappa(n - n_j(n))/n_j(n)$ because $\mu_k \in [0,1]$. Also $U_n \le W_n$, so

$$\frac{1 + U_n(z_j)}{W_n(z_j)^2} \le \frac{1}{n_j(n)} + \frac{1}{n_j(n)^2}.$$

Apply Lemma 9 with $\mathcal{Z} = \mathcal{X}$ to the noise term in the NW decomposition. Summing the resulting error bounds with weights p_j gives the integrated bound. □

Corollary 11 (An explicit count-to-error function)**.** *In this finite model, fix a target pair* (i, j) *throughout the active phase and suppose* $\mathcal{X}_{ij}(u_{ij}) = \{z_1\}$*. For real* $v \in [0, B]$*, set* $b_1(v) = a_1 + v$ *and* $b_j(v) = a_j$ *for* $j > 1$*. Then Assumption 1 holds with the deterministic, finite, nonincreasing function*

$$r_T(v, \eta) = \sum_{j=1}^{m} p_j \left[\kappa\left(\frac{T}{b_j(v)} - 1\right) + \nu\sqrt{2\Lambda_m\left(\frac{1}{b_j(v)} + \frac{1}{b_j(v)^2}\right)}\right]. \tag{8}$$

Proof of Corollary 11. The counts satisfy $n_1(T) = a_1 + M_B$ and $n_j(T) \ge a_j$ for $j > 1$. Each term of $f_T(v)$ is nonincreasing for $0 < v \le T$. Proposition 10 therefore gives $\max_k e_k \le \sum_j p_j f_T(b_j(M_B)) = r_T(M_B, \eta)$ on an event of probability at least $1 - \eta$. The same monotonicity proves that r_T is nonincreasing. □

The diagnostic type's noise contribution decreases as $(a_1 + M_B)^{-1/2}$, up to logarithmic factors; the other types remain controlled by warmup. A diagnostic region containing several types requires control of their individual visit counts.

C.3 Uniform estimation for anchored Gaussian generation

We now use the simultaneous noise bound to verify the estimation conditions in Proposition 3 and Lemma 2. Use the continuous setting and fixed target pair of Appendix A.2, with Gaussian NW bandwidth h, outcomes in $[0,1]$, and Assumption 2. Suppose every μ_k is L-Lipschitz on $[0,1]$, with $L \ge 0$, and fix $\eta \in (0,1)$. As in Appendix C.2, n counts all evaluations, including warmup. Choose a deterministic grid $\mathcal{Z}$ of size G such that every point in $[0,1]$ is within ω of the grid. Define $b_n(z) = \sum_{s=1}^{n} A_{n,s}(z)|x_s - z|$. On the event of Lemma 9, for all k, n,

$$\sup_{x\in[0,1]} |\hat{\mu}_{k,n}(x) - \mu_k(x)| \le E_n, \tag{9}$$

$$E_n := \max_{z\in\mathcal{Z}} \left\{ L b_n(z) + \nu \frac{\sqrt{2\Lambda_G(\eta)(1 + U_n(z))}}{W_n(z)} \right\} + \left(L + \frac{1}{4h^2}\right)\omega.$$

Derivation of equation 9. The bias on the grid is at most $Lb_n(z)$, and Lemma 9 bounds the noise term there. Differentiating the NW estimate gives

$$\frac{d}{dx}\hat{\mu}_{k,n}(x) = \frac{\mathrm{Cov}_{A_n(x)}(Y_k(x_s), x_s)}{h^2}.$$

Both variables lie in $[0,1]$, so their weighted covariance has magnitude at most $1/4$. The estimate is therefore $1/(4h^2)$-Lipschitz. Combining this bound with Lipschitz continuity of μ_k extends the grid bound to all x, including adaptive anchors and candidates. □

To obtain a deterministic bound, divide $[0,1]$ into $J_0 \geq 1$ equal bins, where J_0 is an integer fixed before evaluation. Independent uniform warmup places at least $T_0/(2J_0)$ queries in each bin except on an event of probability at most $J_0 e^{-T_0/(8J_0)}$, by a multiplicative Chernoff bound and a union bound. When every bin meets this count, every $x \in [0,1]$ shares a bin with at least this many warmup observations. Hence, for $T_0 \leq n \leq T$,

$$W_n(x) \geq w_0 := \frac{T_0}{2J_0} e^{-1/(2J_0^2 h^2)}.$$

For any fixed $R \in (0,1]$, splitting observations at distance R gives $b_n(x) \leq R + Te^{-R^2/(2h^2)}/w_0$. Since $U_n \leq W_n$, the right-hand side of equation 9 is at most

$$E_* := L\left(R + \frac{Te^{-R^2/(2h^2)}}{w_0}\right) + \nu\sqrt{2\Lambda_G(\eta)\left(\frac{1}{w_0} + \frac{1}{w_0^2}\right)} + \left(L + \frac{1}{4h^2}\right)\omega. \tag{10}$$

Corollary 12 (Diagnostic acquisition for anchored Gaussian generation)**.** *Under the preceding conditions and the warmup and anchor conditions of Proposition 3, let $\beta_{\max} = \max_{t \leq B} \beta_t$ and let p_* be given by equation 4. If*

$$4E_* < \zeta, \qquad 4E_* + \frac{\beta_{\max}}{\sigma\sqrt{w_0}} \leq \gamma - u,$$

then, for any $\delta_c \in (0,1)$,

$$\mathbb{P}\left(M_B \geq p_* B - \sqrt{\frac{B}{2}\log\frac{1}{\delta_c}}\right) \geq 1 - \eta - \delta_c - (1-s)^{T_0} - J_0 e^{-T_0/(8J_0)}. \tag{11}$$

Proof of Corollary 12. Intersect the simultaneous noise event, the warmup bin-coverage event, and the event that warmup hits S. Their total failure probability is at most $\eta + J_0 e^{-T_0/(8J_0)} + (1-s)^{T_0}$. On this intersection, Proposition 3 ensures a diagnostic anchor in every round, so equation 4 gives $Q_B \geq p_* B$. Lemma 2 gives $L_B = 0$. Intersect with the unconditional event in Theorem 1, whose probability is at least $1 - \delta_c$, and apply a union bound. No conditioning of the martingale inequality on future estimation accuracy is required. □

The corollary concerns a fixed target pair and exact NW estimates; its conditions may be conservative at small budgets and do not certify the experimental budgets. For a numerical implementation that replaces NW predictions when kernel mass is very small, the result applies when w_0 exceeds that replacement threshold.

C.4 Correct selection in a finite query model

We combine the acquisition and estimation bounds in the finite model of Appendix C.2, using a generator that samples candidates from an anchor list and from $\mathcal{D}$. Fix one target pair (i,j) throughout the active phase, let $d_\ell = |D_{ij}(z_\ell)|$, and write $\gamma = \gamma_{ij}$ and $u = u_{ij}$. Assume

$$d_1 \geq \gamma > u > \max_{\ell > 1} d_\ell, \qquad \zeta = d_1 - \max_{\ell > 1} d_\ell > 0.$$

Thus z_1 is the only diagnostic type at either threshold. At each round, the anchor list contains at most $A_{\max}$ evaluated observations and includes a maximizer of the estimated absolute pair gap. Each of N conditionally independent candidates is drawn uniformly from this list with probability $\lambda \in [0,1]$, and from $\mathcal{D}$ otherwise. The generator does not require knowledge of z_1. A fixed candidate pool may also be included.

Fix $\eta, \delta_c \in (0,1)$, with $\Lambda_m = \Lambda_m(\eta)$ as in Appendix C.2. Let $a_{\min} = \min_j a_j$ and $\beta_{\max} = \max_{t \leq B} \beta_t$, and define

$$E_{\text{fin}} := \frac{\kappa T}{a_{\min}} + \nu\sqrt{2\Lambda_m\left(\frac{1}{a_{\min}} + \frac{1}{a_{\min}^2}\right)}, \tag{12}$$

$$p_* := 1 - \left(1 - \frac{\lambda}{A_{\max}} - (1-\lambda)p_1\right)^N, \qquad n_0 := \max\left\{0, p_* B - \sqrt{\frac{B}{2}\log\frac{1}{\delta_c}}\right\}. \tag{13}$$

Theorem 13 (Selection with explicit coverage and estimation bounds)**.** *In this finite model, let r_T be given by equation 8. If*

$$4E_{\text{fin}} < \zeta, \qquad 4E_{\text{fin}} + \frac{\beta_{\max}}{\sigma\sqrt{a_{\min}}} \le \gamma - u, \qquad r_T(n_0, \eta) < \Delta_{\min}/2, \tag{14}$$

then $\mathbb{P}(\hat{k} = k^*) \ge 1 - \eta - \delta_c$.

Proof of Theorem 13. Let $\mathcal{E}$ be the event in Proposition 10, which has probability at least $1 - \eta$. Because $n_j(n) \ge a_{\min}$ and $n \le T$, all conditional-value estimates on this event have absolute error at most E_{fin} throughout the active phase. Warmup includes z_1, whose estimated absolute gap exceeds that of every other type by at least $\zeta - 4E_{\text{fin}} > 0$. The anchor list therefore contains z_1 in every round. Each generated candidate hits z_1 with conditional probability at least $\lambda/A_{\max} + (1 - \lambda)p_1$, so $q_t \ge p_*$ and $Q_B \ge p_* B$ on $\mathcal{E}$. Moreover, $W_n(z_j) \ge a_{\min}$, and the second condition in equation 14 gives $L_B = 0$ by Lemma 2.

Intersect $\mathcal{E}$ with the event in Theorem 1, whose probability is at least $1 - \delta_c$. On this intersection, $M_B \ge n_0$. The error bound of Corollary 11 already holds on $\mathcal{E}$, so $\max_k e_k \le r_T(M_B, \eta) \le r_T(n_0, \eta) < \Delta_{\min}/2$. Proposition 5 gives correct selection. The common estimation event is counted once, giving total failure probability at most $\eta + \delta_c$. □

Theorem 13 and Corollary 12 both fix the target pair. For $K = 2$, the unordered pair is fixed automatically; for $K > 2$, a changing pair requires additional control of pairwise coverage or pair selection. Neither result establishes higher selection accuracy than random sampling.

D Additional Experimental Results

This appendix provides implementation details and results for the additional pool configurations and budgets in Section 5.

D.1 Experimental settings

Common settings. All three studies use $\sigma = 0.5$ in the uncertainty proxy. The warmup budget is $T_0 = \min\{\max(4K, \lfloor T/4 \rfloor), \lfloor T/2 \rfloor\}$, and the uncertainty-bonus coefficient is $\beta_t = C_0\sqrt{2\log(2Kt^2/\delta)}$, with $C_0 = 0.2$ and $\delta = 0.05$. The number of generated candidates, N_{gen}, corresponds to N in Section 3. DASE Pool Only sets $N_{\text{gen}} = 0$, and DASE Gen Only sets $\mathcal{C}_{\text{pool}} = \emptyset$. TradQBC uses the variance of $\hat{\mu}_k(x)$ across agents, without targeting a particular pair or adding an uncertainty bonus.

Synthetic study. At each active round, the generator uses up to five evaluated queries with the largest estimated leader–challenger gaps, $|\hat{\mu}_i(x_s) - \hat{\mu}_j(x_s)|$, as anchors. It draws 15 candidates from Gaussian distributions centered at these anchors, with standard deviation h and truncated to $[0, 1]$. Five additional candidates are sampled uniformly from $[0, 1]$ to explore the full query space. The central interval $[0.47, 0.53]$ occupies 6% of the query space; Restricted Pool excludes grid points in this interval.

MATH study. The query subset is sampled from the training split with stratification by subject and difficulty. The test split is used to compute deployment values and the averages $\hat{V}_k$ and $\hat{S}_k$ and is never queried. The query subset approximates its subject and difficulty composition and serves as the sampling population for warmup and Random. The feature vector consists of difficulty, seven binary subject indicators (a one-hot encoding), normalized reference-solution length, and binary indicators of computational and proof content in the reference solution. These features assume access to reference solutions before query selection. Responses are graded using symbolic equality in SymPy (Meurer et al., 2017), with string matching as a fallback (Appendix E). Each evaluation returns the recorded binary outcome, independently flipped with probability $\epsilon = 0.01$. In the MATH and clinical studies, let $s_k(x) \in \{0, 1\}$ denote the recorded score. The observed outcome has mean $\mu_k(x) = \epsilon + (1 - 2\epsilon)s_k(x)$, giving $V_k = \epsilon + (1 - 2\epsilon)\mathbb{E}_{X\sim\mathcal{D}}[s_k(X)]$. The reported deployment

values include this noise, which preserves the agent ranking and scales all value gaps by 0.98. Fixed-pool candidates and Random's queries exclude problems already evaluated in the trial. Generated candidates are not filtered in this way, so a generated problem may be evaluated more than once. Warmup queries are stratified by subject, with every subject represented. Random samples active-phase queries uniformly from the query subset, independently of the fixed pool.

Clinical study. The dataset contains 2,400 admissions, each with the history of present illness, physical examination, laboratory tests with reference ranges, microbiology, and radiology reports. Both language models are 4-bit quantized and served locally on institution-owned hardware, following PhysioNet's recommendation (PhysioNet, 2025). Appendix F details the evaluation procedure, sampling, and retrieval generator. On D, the recorded outcomes of the top two agents differ on 23.8% of admissions. Their recorded diagnosis accuracies differ by 6.4 percentage points within cholecystitis; each of the other three pathologies contributes between -0.4 and 0.4 percentage points to the overall accuracy margin.

D.2 Synthetic study

Table 3 reports results with standard errors for all three pools, including the Standard Pool results summarized in Table 1. **Standard Pool** is a 50-point uniform grid over $[0, 1]$. **Restricted Pool** removes the points in the central interval $[0.47, 0.53]$ from Standard Pool; the remaining points can still have nonzero pairwise gaps. **Small Pool** is a 10-point uniform grid whose two points nearest the center are 0.455 and 0.545. Like Restricted Pool, this pool has no points in the central interval. Random and DASE Gen Only do not use the fixed pool, so their rows are identical across panels.

Under Restricted Pool, DASE Gen Only and DASE Full exceed both pool-restricted methods at every tested budget. The same pattern holds under Small Pool.

Table 3: Best-agent selection accuracy in the synthetic study for Standard, Restricted, and Small pools, estimated from 1,000 independent trials per pool and budget. Each T includes warmup. Entries are mean $\pm$ standard error over the trials; bold marks the highest point estimate, not statistical significance.

Method	$T = 100$	$T = 200$	$T = 300$	$T = 500$
Standard				
Random	0.559±0.016	0.624±0.015	0.715±0.014	0.801±0.013
TradQBC	0.498±0.016	0.565±0.016	0.587±0.016	0.643±0.015
DASE Pool Only	**0.596±0.016**	0.711±0.014	0.830±0.012	0.925±0.008
DASE Gen Only	0.568±0.016	**0.738±0.014**	**0.845±0.011**	**0.932±0.008**
DASE Full	0.567±0.016	0.709±0.014	0.805±0.013	0.930±0.008
Restricted				
Random	0.559±0.016	0.624±0.015	0.715±0.014	0.801±0.013
TradQBC	0.453±0.016	0.497±0.016	0.539±0.016	0.597±0.016
DASE Pool Only	0.507±0.016	0.570±0.016	0.611±0.015	0.710±0.014
DASE Gen Only	**0.568±0.016**	**0.738±0.014**	**0.845±0.011**	**0.932±0.008**
DASE Full	0.555±0.016	0.673±0.015	0.795±0.013	0.909±0.009
Small				
Random	0.559±0.016	0.624±0.015	0.715±0.014	0.801±0.013
TradQBC	0.495±0.016	0.518±0.016	0.575±0.016	0.627±0.015
DASE Pool Only	0.547±0.016	0.619±0.015	0.695±0.015	0.786±0.013
DASE Gen Only	**0.568±0.016**	**0.738±0.014**	**0.845±0.011**	**0.932±0.008**
DASE Full	0.553±0.016	0.675±0.015	0.810±0.012	0.925±0.008

With Standard Pool at $T = 500$, TradQBC allocates 87% of its active-phase queries to points within 0.03 of $x = 0.15$ or $x = 0.85$, and 11% to the central interval $[0.47, 0.53]$. The corresponding proportions are 6.5% and 48% for DASE Pool Only, 6% and 46% for both DASE Gen Only and DASE Full, and 12% and 6% for Random.

D.3 MATH STUDY

Table 4 describes the three fixed pools. Table 5 reports results with standard errors for all three pools, including the Standard Pool results summarized in Table 1. **Standard Pool** and **Small Pool** are stratified samples of 500 and 100 problems from the query subset. **Restricted Pool** is a separate stratified sample of 500 problems drawn from the query subset after excluding Number Theory. It therefore contains no diagnostic problem for the constructed comparison between agents 4 and 5. The generator retrieves candidates from the 5,000-problem generation pool in all three configurations.

Under Restricted Pool, both generation variants exceed TradQBC and DASE Pool Only at every tested budget.

Under Small Pool, DASE Pool Only leads at $T = 75$ and $T = 100$, but both pool-restricted methods exhaust the available queries before reaching $T = 200$. DASE Gen Only has the highest accuracy at this budget.

Table 4: Fixed candidate pools in the MATH study. Size is the total number of problems; Number Theory is the number in that subject. Diagnostic problems have different recorded outcomes for Self-Consistency ($k = 4$) and NT-Diagnostic ($k = 5$).

Configuration	Size	Number Theory	Diagnostic problems
Standard	500	58	6
Restricted	500	0	0
Small	100	12	1

Table 5: Best-agent selection accuracy in the MATH study for Standard, Restricted, and Small pools, estimated from 500 independent trials per pool and budget. Each T includes warmup. Random and DASE Gen Only do not use the fixed pool, so their rows are identical across panels. Entries are mean $\pm$ standard error over the trials; bold marks the highest point estimate, not statistical significance.

Method	$T = 50$	$T = 75$	$T = 100$	$T = 200$
Standard				
Random	0.592±0.022	0.674±0.021	0.722±0.020	0.838±0.016
TradQBC	0.602±0.022	0.656±0.021	0.686±0.021	0.760±0.019
DASE Pool Only	**0.660±0.021**	**0.756±0.019**	**0.826±0.017**	0.896±0.014
DASE Gen Only	0.646±0.021	0.702±0.020	0.760±0.019	0.844±0.016
DASE Full	0.628±0.022	0.754±0.019	0.810±0.018	**0.900±0.013**
Restricted				
Random	0.592±0.022	0.674±0.021	0.722±0.020	0.838±0.016
TradQBC	0.564±0.022	0.598±0.022	0.622±0.022	0.708±0.020
DASE Pool Only	0.582±0.022	0.578±0.022	0.624±0.022	0.720±0.020
DASE Gen Only	0.646±0.021	0.702±0.020	0.760±0.019	0.844±0.016
DASE Full	**0.666±0.021**	**0.714±0.020**	**0.774±0.019**	**0.858±0.016**
Small				
Random	0.592±0.022	0.674±0.021	0.722±0.020	0.838±0.016
TradQBC	0.616±0.022	0.696±0.021	0.718±0.020	0.818±0.017
DASE Pool Only	0.616±0.022	**0.796±0.018**	**0.792±0.018**	0.802±0.018
DASE Gen Only	**0.646±0.021**	0.702±0.020	0.760±0.019	**0.844±0.016**
DASE Full	0.590±0.022	0.682±0.021	0.754±0.019	0.828±0.017

D.4 UNION RANDOM BASELINE

To assess the effect of access to the generation pool, we compare Random with **Union Random**. Union Random uses the same warmup queries and then samples uniformly without replacement from the unevaluated queries in the union of the query subset and the generation pool. It never queries the reference set, and the union is not exhausted at the tested budgets. Table 6 compares the

two methods on the same trials. Neither method depends on the fixed candidate pool, so their results are identical across Standard, Restricted, and Small settings.

Union Random has lower accuracy than Random on MATH at all tested budgets and similar accuracy in the clinical study. Thus, access to the additional queries alone does not consistently improve selection accuracy.

Table 6: Random and Union Random: best-agent selection accuracy (mean ± standard error) from 500 trials per cell on MATH and 1,000 in the clinical study. Both methods share the trials and warmup queries of the main experiments. Results apply to all three fixed-pool settings.

	MATH				Clinical		
Method	$T = 50$	$T = 75$	$T = 100$	$T = 200$	$T = 90$	$T = 120$	$T = 150$
Random	0.592±0.022	0.674±0.021	0.722±0.020	0.838±0.016	0.327±0.015	0.340±0.015	0.369±0.015
Union Random	0.550±0.022	0.666±0.021	0.704±0.020	0.786±0.018	0.323±0.015	0.353±0.015	0.366±0.015

E MATH STUDY: PROMPTS AND DECODING

This appendix gives the four LLM agents' prompts verbatim, together with the decoding and grading procedures used in the MATH study (Section 5.2). The fifth agent, NT-Diagnostic, uses the recorded outcomes of Self-Consistency outside Number Theory and Zero-shot CoT on Number Theory problems; it makes no additional model calls.

E.1 AGENT PROMPTS

All four agents use GPT-4o-mini in the same evaluation framework, with the prompts and decoding settings specified below. The placeholder ⟨PROBLEM⟩ denotes the MATH problem statement, inserted verbatim without preprocessing.

$k = 1$ **(Direct).** System message:

> *You are a math solver. Give your final answer inside* `\boxed{}`.

User message: ⟨PROBLEM⟩.

$k = 2$ **(Zero-shot CoT, Kojima et al., 2022).** System message:

> *You are a math solver.*

User message:

> ⟨PROBLEM⟩
> *Let's think step by step. Put your final answer inside* `\boxed{}` *at the end.*

$k = 3$ **(Four-shot CoT, Wei et al., 2022).** System message:

> *You are a math solver.*

The user message contains the four worked examples below, followed by "*Now solve:*", the problem statement, and "*Solution:*". The examples are fixed across problems and cover four of the seven MATH subjects: Algebra, Geometry, Number Theory, and Counting & Probability.

> *Example 1 (Algebra):*
> *Problem: Solve for x: 2x + 3 = 11.*
> *Solution: Subtracting 3 from both sides gives 2x = 8, so x = 4.* `\boxed{4}`
> *Example 2 (Geometry):*
> *Problem: A right triangle has legs of length 3 and 4. What is its hypotenuse?*

Solution: By the Pythagorean theorem, hypotenuse = `\sqrt{3^2+4^2} = \sqrt{25} = 5. \boxed{5}`
Example 3 (Number Theory):
Problem: What is the remainder when 17 is divided by 5?
*Solution: 17 = 3*5 + 2, so the remainder is 2.* `\boxed{2}`
Example 4 (Counting & Probability):
Problem: How many ways can you arrange the letters in the word CAT?
Solution: There are 3! = 6 arrangements. `\boxed{6}`

$k = 4$ **(Self-Consistency, Wang et al., 2023).** Self-Consistency uses the same messages as Zero-shot CoT and samples 10 completions at temperature 0.7. It extracts the boxed answer from each completion and returns the majority answer among successful extractions. If no completion contains a boxed answer, the procedure falls back to the first completion's extraction, which is scored as incorrect.

E.2 Decoding and grading

All agents use GPT-4o-mini with a 1024-token completion limit. Direct, Zero-shot CoT, and Four-shot CoT use greedy decoding at temperature 0, subject to any nondeterminism in the provider's implementation. Self-Consistency uses the sampling procedure described above.

The grader extracts the final boxed expression from both the completion and the reference solution, parses them with SymPy's LaTeX parser, and marks them as equal if their difference simplifies to zero (Meurer et al., 2017). If parsing fails, it compares strings after removing whitespace and dollar signs and ignoring case. A completion without a boxed answer receives a score of 0.

The grader is deterministic, and the $5 \times 12{,}500$ outcome matrix is computed once. All subsequent trials retrieve recorded outcomes from this matrix. During sequential evaluation, each retrieved outcome is flipped independently with probability $\epsilon = 0.01$ (Appendix D.1).

F Clinical Study: Implementation Details

This appendix provides implementation details for the clinical study (Section 5.3), including the evaluation procedure, agent protocols, diagnosis scoring, sampling, query features, and generation. It also evaluates an LLM-based implementation that proposes queries from retrieved candidates.

F.1 Evaluation procedure

For each admission, an agent first receives the history of present illness. It then requests additional information, such as physical examination findings, laboratory results, or radiology reports. The evaluation framework returns the corresponding information from the admission record, and the agent continues until it provides a final diagnosis. This complete interaction constitutes one rollout. For each admission and agent, we record a binary diagnosis score: 1 if the final diagnosis matches the recorded pathology under the deterministic evaluator of Hager et al. (2024), and 0 otherwise (Appendix F.3).

We store these diagnosis scores in a matrix indexed by admission and agent. In each selection round, the method chooses one admission and observes an outcome $Y_k(x)$ for each of the six agents. This outcome is the stored diagnosis score independently flipped with probability 0.01, as in the MATH study. The top two agents have deployment values of approximately 0.637 and 0.620, with a gap of approximately 0.018 after accounting for this noise (Appendix D.1). Queries retrieved by the generator are evaluated in the same way, so the selection experiments require no additional rollouts.

F.2 Retrieval protocols

All three protocols use the evaluation framework of Hager et al. (2024) with the same decoding settings. They differ only in the system instruction appended to the shared prompt. P0 adds no instruction. P1 and P2 append the text below verbatim, with HPI denoting the history of present illness.

P1 (ranked differential diagnosis).

> *Before requesting any information beyond the HPI, first state your differential diagnosis as a ranked list of possible conditions. For each subsequent information request, briefly state which item(s) on your differential it is meant to help you rule in or out. Update your differential after each new piece of information you receive.*

P2 (budget-constrained).

> *You have a maximum of 5 total information requests for this case. Use them judiciously.*

These instructions modify how the agent requests and organizes information without naming a specific pathology or diagnostic test.

F.3 Diagnosis scoring

Free-text diagnoses are scored using the three deterministic steps of Hager et al. (2024), without modification or a language model. First, pattern matching extracts the diagnosis from the final message using phrases such as "the diagnosis is X" or "the patient has X". If several conditions are named, only the first is retained because the evaluation requires a single diagnosis. Second, the evaluator compares the extracted string with the admission's pathology label from the dataset. It accepts either a fuzzy match to the pathology name or a match to a curated list of alternative terms requiring an anatomical location and a qualifying modifier. For example, "gallbladder inflammation" matches cholecystitis, whereas "gallbladder" alone does not. Third, each match must pass a negation check, which rejects phrases such as "no evidence of appendicitis". We use the more permissive diagnosis score provided by Hager et al. (2024). The recorded score is binary and reproducible, although the matching rules can reject a clinically reasonable answer with unrecognized wording.

F.4 Experimental protocol

Table 2 reports all five methods. The reference set D contains 125 admissions per pathology and is the same in every trial. The remaining 399 admissions (99 with appendicitis and 100 with each other pathology) are divided anew in each trial. The query subset contains 50 admissions per pathology, and the remaining 199 form the generation pool. Standard Pool is the query subset, Small Pool contains 25 admissions per pathology sampled from it, and Restricted Pool contains its 150 admissions remaining after excluding cholecystitis. Within trial s, all methods and pool configurations share the division and warmup queries, yielding paired comparisons.

All methods, including Random, draw warmup queries from the query subset with stratification by pathology and at least one admission per class. The warmup budget is $\min\{\max(4K, \lfloor T/4 \rfloor), \lfloor T/2 \rfloor\}$, giving 24, 30, and 37 queries at the three reported budgets. During the active phase, Random samples uniformly from the query subset, TradQBC and DASE Pool Only select from the fixed pool, and DASE Gen Only selects among generated candidates. DASE Full selects from the union of generated candidates and the N_{gen} pool admissions with the highest acquisition scores. Each admission is queried at most once within a trial; outcomes follow the procedure in Appendix F.1. The reported budgets are $T \in \{90, 120, 150\}$, including warmup, with $1{,}000$ trials per cell. At $T = 120$ and $T = 150$, the active phase allows 90 and 113 queries. Under Small Pool, the pool-restricted methods exhaust the unevaluated pool admissions and stop early in most trials at $T = 120$ and in all trials at $T = 150$.

We redraw the division to account for variation in pool composition. The top two agents' recorded diagnosis accuracies differ by 0.018 on D. Their accuracy difference on a random 200-admission subset has a standard deviation of about 0.03.

F.5 Feature vector and normalization

Each admission is represented by $\varphi(x) \in \mathbb{R}^9$, comprising four binary indicators for the known pathology, the comorbidity count, an indicator of whether a procedure was performed, the fraction of laboratory values outside their reference ranges, the number of radiology reports, and the word

count of the history of present illness. These features use structured records and counts, without free-text content. The pathology indicators use dataset labels, so this evaluation assumes that the query-selection procedure has access to those labels.

We normalize each of the three count features by its 95th percentile in the pool being used and clip the result to $[0, 1]$. The percentile is recomputed for each pool, without using the full set of $2{,}400$ admissions.

F.6 Retrieval generator

Candidate batch. Each active round forms a batch of at most $N_{\text{gen}} = 50$ candidates from the generation-pool admissions not yet evaluated in the trial. The generator computes NW estimates at the evaluated admissions and uses the five with the largest estimated gap $|\hat{\mu}_i(x) - \hat{\mu}_j(x)|$ for the current leader–challenger pair as anchors. For a full batch, each anchor contributes its seven nearest admissions in the generation pool, measured by squared Euclidean distance on φ. This gives 35 local candidates, supplemented by 15 sampled uniformly from the generation pool. The same acquisition function scores generated and fixed-pool candidates. We exclude previously evaluated admissions to avoid retrieving the same cases repeatedly.

Generation-pool coverage. The generation pool contains approximately 50 admissions per pathology. Under Restricted Pool, it is the only source of cholecystitis admissions during the active phase, limiting their number to at most 50 even at $T = 150$. The clinical generation pool is comparable in size to Standard Pool, whereas the MATH generation pool contains $5{,}000$ problems, ten times the size of Standard Pool.

F.7 Language-model selection among retrieved candidates

We evaluate LLM-guided query proposal using a locally served Qwen2.5-32B model (Qwen Team, 2025). The model proposes queries from retrieved candidates with recorded outcomes, enabling evaluation without annotating new clinical cases. We denote these variants by **DASE Gen Only (LLM)** and **DASE Full (LLM)**. Given the leader–challenger pair, evaluation history, and excerpts from the ten retrieved admissions with the highest acquisition scores, the model selects three admissions for the next three rounds that use generated queries. The retrieval procedure and acquisition ranking are unchanged. DASE Full (LLM) invokes the model only when the highest-scoring generated candidate is at least as highly ranked as the best fixed-pool candidate. Table 7 compares these variants with Random, TradQBC, and DASE Pool Only on the same 200 trials. DASE Full (LLM) exceeds TradQBC and DASE Pool Only under all three pools (0.420, 0.390, and 0.320 against at most 0.410, 0.370, and 0.245) and Random under Standard and Small pools; under Restricted Pool it is within one standard error of Random (0.320 against 0.340). DASE Gen Only (LLM) matches Random (0.340). Using cases with recorded outcomes makes LLM-guided query proposal directly evaluable, while restricting the model to a retrieved shortlist. Performance therefore depends on both the shortlist's coverage and the local model's ability to identify diagnostic queries.

Table 7: Best-agent selection accuracy of LLM variants, Random, TradQBC, and DASE Pool Only at $T = 120$, using the same 200 trials per pool configuration. Entries are mean $\pm$ standard error over the trials; bold marks the highest point estimate, not statistical significance.

Method	Standard	Small	Restricted
Random	0.340±0.033	0.340±0.033	**0.340±0.033**
TradQBC	0.375±0.034	0.370±0.034	0.245±0.030
DASE Pool Only	0.410±0.035	0.355±0.034	0.245±0.030
DASE Gen Only (LLM)	0.340±0.033	0.340±0.033	**0.340±0.033**
DASE Full (LLM)	**0.420±0.035**	**0.390±0.034**	0.320±0.033