SKETCH-INFORMED CAPABILITY DISCOVERY IN LARGE AGENT COLLECTIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

Open collectives of foundation models, agents, and web services now contain thousands of entities, and a coordinator who needs a qualified entity for a task cannot afford to inspect them all. We study the coordinator's discovery problem under a two-tier query model, where each of the K entity-groups publishes a free summary of its members' capabilities, alongside a costly query that reveals a random entity from a chosen group. Our main result is a threshold on the summary dimension with resolution β. When the dimension is $O(\beta^{-2} \log K)$, summaries are informative enough to guide the search. Our proposed SIEVE algorithm uses the summaries to eliminate groups and then queries the remaining groups' members. With probability $1 - \delta$, SIEVE discovers and returns an entity that can perform the task and ranks in the top q fraction of the collective, counted over the groups that contain the capability, with at most $O(q^{-1} \log \delta^{-1})$ queries. This query complexity depends on neither K nor the size of the collective, and no algorithm with a worst-case query limit can improve it by more than a constant factor. Below the dimension threshold, we prove in a Gaussian model of the summary readout that the query complexity of any coordinator that reads the summaries through inner products grows polynomially in K. Experiments on synthetic populations with up to one million entities, as well as on RapidAPI and the Open LLM Leaderboard, support the predicted dimension threshold and query savings.

1 INTRODUCTION

Multi-agent systems are expanding beyond small, isolated teams into open collectives, bringing together thousands of entities, from foundation models to programming tools and web APIs, each contributing a fragment of work to solve complex tasks (Tran et al., 2025; Shen et al., 2023; Chen et al., 2025; Qin et al., 2024). Coordination protocols specify entities interaction and invoke tools once connected, while finding the right entity among thousands of available services is left to a discovery and registry layer (Surapaneni et al., 2025; Hou et al., 2026; Muscariello et al., 2025). Each entity publishes a profile of its constraints and capabilities, and a coordinator matches the requirements of a task against these profiles to find a suitable one (Guo et al., 2026; Tokmak et al., 2024; Klusch et al., 2016; Meshkova et al., 2008). The cost of this search grows with the size of the collective and with every capability update. When the capability is rare, the coordinator inspects many profiles before it finds one that qualifies, even if the collective is organized.

We move the discovery burden from individuals to groups. Each group publishes a compact summary of the combined capability of its members, and the coordinator filters at the group level before querying individuals. Two-tier designs of this kind appear in pooled testing and spectrum sensing (Aldridge et al., 2019; Jin et al., 2012), and we adopt them for capability matching in multi-agent systems. In the coalition formation literature, a coalition aggregates its members' capabilities into a vector to test whether a task is feasible for that team (Doostmohammadian et al., 2025). Our focus is the complementary setting where a coordinator faces many teams that have already published their summaries. We build group's summary as a fixed-dimension sketch by summing random vector embeddings assigned to its members' capability labels (Rahimi & Recht, 2007; Wacker et al., 2023). Because the summary is additive, a group member add or removal requires a local update instead of rebuilding the sketch from scratch (Cormode & Muthukrishnan, 2005; Agarwal et al., 2013).

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.

In this paper, we look for the minimum summary dimension that helps with the coordinator's search. With K groups, say a group qualifies if at least β fraction of its members qualify. The threshold dimension is of order $\beta^{-2} \log K$. Above it, the summaries show which group contains the capability. Below it, noise makes a group without the capability appear capable, and the summaries mislead the coordinator. Building on this threshold, our contributions are as follows:

- **A model of two-tier discovery.** We formalize discovery over a registry in which every group publishes a summary, an additive sketch of dimension d, and every query reveals one random member of a chosen group. Success means returning an entity that can perform the task and ranks in the top q fraction of the collective counted over the groups that contain the capability, and we call a summary dimension decision sufficient when it keeps the number of queries within a given budget (Sec. 3).
- **An algorithm whose query complexity is independent of the collective.** SIEVE uses the summaries to eliminate groups, draws members from the remaining groups in proportion to their estimated number of qualified members, and checks each revealed profile against the task (Sec. 4). With a summary dimension of order $\beta^{-2} \log K$, it succeeds with probability $1-\delta$ using at most $O(q^{-1} \log \delta^{-1})$ queries, independent of K and the size of the collective, and no algorithm with a worst-case query limit can improve this bound by more than a constant factor (Sec. 5).
- **A matching lower bound.** Below that dimension, over a population where a single group contains the capability, any coordinator that reads the summaries through inner products with Gaussian readout noise needs a number of queries that grows polynomially in K (Sec. 5). On the same population, a search without summaries needs $O(K/\beta)$ queries, while SIEVE needs $O(1/\beta)$. The summaries let the coordinator pay for the density of the capability instead of the number of groups.
- **Experiments.** On synthetic populations with up to one million entities and two public registries, the RapidAPI corpus of over 16,000 web APIs and the Open LLM Leaderboard, the number of queries drops at the predicted dimension threshold and the savings match the theory (Sec. 6).

Sec. 2 positions this work in the literature, Sec. 7 concludes, and the Appendix contains complete proofs.

2 RELATED WORK

Maximizing team performance drives the multi-agent systems literature toward improving candidate selection, group composition, and communication. Recent orchestration frameworks treat team formation as an online search or sampling problem (Zhuge et al., 2024; Zhang et al., 2025b;a), score agents via trial runs (Liu et al., 2023), or redesign interaction graphs (Wang et al., 2025; Qian et al., 2025), demonstrating gains over single agent problem solving and competing team designs. These methods presume a closed candidate set that can be directly audited or trial-queried.

Beyond a single team formation, locating capable entities across heterogeneous groups shares foundations with multi-arm bandits and pure exploration. Candidate entities correspond to arms, and discovery means identifying a "good arm" whose performance exceeds a certain threshold (Even-Dar et al., 2002; Kano et al., 2019; Even-Dar et al., 2006; Mannor & Tsitsiklis, 2004). For large populations, infinite-arm and reservoir bandits model the setting where an algorithm cannot test every candidate and must declare a near-optimal option from an unknown population (Berry et al., 1997; Carpentier & Valko, 2015; Katz-Samuels & Jamieson, 2020). In grouped max-quantile exploration, arms are in K distinct groups (reservoirs) and the goal is to identify the group with the highest reward quantile (Lau et al., 2023), while quantile pure exploration over a single reservoir seeks one arm in a top fraction (Gong & Sellke, 2023). Such methods estimate group quality by sampling $\Omega(\epsilon^{-2} \log(K/\delta))$ arms per group, where δ is the failure probability and ϵ is the target accuracy threshold.

Our two-tier query model is related to multi-fidelity and memory-constrained exploration. Multi-fidelity bandits split a query budget between cheap but imperfect evaluations and expensive accurate observations (Kandasamy et al., 2016; Poiani et al., 2022; 2024). The low-fidelity channel is analogous to our group-sketch registry (Assum. 2). Cost-aware LLM routing employs a similar principle,

using inexpensive models before invoking frontier APIs (Chen et al., 2023). Memory-limited and streaming bandits restrict the number of individual arm identities an algorithm can store while tracking optimal candidates in an input stream (Assadi & Wang, 2020; Maiti et al., 2021; Assadi & Wang, 2024). Our memory holds one summary per group, and the constraint is on the dimension of that summary.

The memory representation of our group summaries is rooted in streaming sketches, kernel embeddings, and vector-symbolic architectures. Mergeable structures such as count-min sketches (Cormode & Muthukrishnan, 2005; Agarwal et al., 2013) and Bloom filters (Bloom, 1970; Broder & Mitzenmacher, 2004) summarize the capability labels of members of a group. For example, a count-min sketch can estimate how many members have a known capability, while a Bloom filter can test whether at least one member has it. Summaries from separate subgroups can then be combined efficiently (Cormode & Muthukrishnan, 2005; Agarwal et al., 2013). For richer continuous attributes, random Fourier features and phasor embeddings on the complex torus provide additive, unbiased kernel representations (Rahimi & Recht, 2007; Plate, 1995; Kleyko et al., 2022). Prior work studies how such sketches can reconstruct data or decode superposed representations (Keriven et al., 2018; Gribonval et al., 2017; Clarkson et al., 2026), but our focus is to determine the smallest sketch dimension that can reliably guide downstream selection and prevent unnecessary individual queries.

Positioning of Our Work: Classical pure exploration and good-arm identification (Even-Dar et al., 2002; Mannor & Tsitsiklis, 2004; Kaufmann et al., 2016; Garivier & Kaufmann, 2016) observe every candidate through a single channel of noisy pulls, and sketch capacity results (Clarkson et al., 2026) bound the dimension needed to read individual items out of a superposition. We split the observation into two levels. A free summary per group and an exact query per member, then ask how many summary dimensions a group must publish at the upper level to save contacts at the lower level. We formalize this question as the decision sufficiency of a sketch family, give an algorithm that solves it, and prove upper and lower bounds on the dimension that match in order.

3 Problem Formulation

The decision maker in this problem is a coordinator who, among a collective of entities $\{1, 2, \ldots, N\}$, wants to select one to carry out a given task. The entities are agents, modules, or APIs, differing in their capabilities, and a task $\tau \in \mathcal{T}$ can be performed only by a subset of them. The coordinator issues queries to identify a capable member whose quality is in the top q capable members of the collective. The N entities are partitioned into K groups, each called a stratum. Stratum c contains n_c members and carries the weight $\pi_c := \frac{n_c}{N}$. Strata may correspond to agents or APIs published under the same provider, model family, or operation region. Each entity i in stratum c holds an attribute record $a_{c,i} \in \mathcal{A}$ encoding properties such as role, tool access, and operating horizon. Two functions act on this record. First, the feasibility function $f_\tau : \mathcal{A} \to \{0, 1\}$ for a given task τ, which indicates whether entity i can perform the task $f_\tau(a_{c,i}) = 1$, or not $f_\tau(a_{c,i}) = 0$. Second, the quality function $q_\tau : \mathcal{A} \to \mathcal{X}_L$, measuring its quality of performance, where $\mathcal{X}_L := \{0, \frac{1}{L}, \frac{2}{L}, \ldots, 1\}$ for a resolution parameter $L \in \mathbb{N}$. Real-valued scores are accommodated by rounding down to the closest bin in $\mathcal{X}_L$ (App. A.2).

For $x \in \mathcal{X}_L$ define the *qualified portion* of stratum c and of the whole collective, respectively, as:

$$T_c(x) := \frac{\left|\{i \in c : f_\tau(a_{c,i}) = 1, q_\tau(a_{c,i}) \geq x\}\right|}{n_c}, \quad T(x) := \sum_{c=1}^{K} \pi_c T_c(x). \tag{1}$$

3.1 Strata Access and Summaries

The coordinator does not directly observe any record $a_{c,i} \in \mathcal{A}$. Instead, it may query a stratum c to reveal an attribute record. The feasibility and quality functions can then determine the entity's qualification for the task.

Assumption 1 (Query Procedure)**.** A direct query to any stratum c reveals the attribute record of a uniformly random member not yet queried in that stratum c. Next draws of the same stratum are without replacement and independent of draws in other strata.

Beyond individual queries, a group-level summary S_c provides a compact representation of the capability distribution across members of stratum c. These summaries are available publicly.

Assumption 2 (Public Registry)**.** The stratum sizes $n_1, \ldots, n_K$, and therefore the weights $\pi_1, \ldots \pi_K$, are available in a public registry. Also each stratum c computes a summary S_c locally and adds it to the registry. The coordinator locally caches or reads $S_1, \ldots, S_K$ at no cost.

Each sketch has a dimension of d and is initially built from the attribute records of all members in the stratum. As time passes, entities may upgrade their capabilities, join a stratum, or leave one. To model these changes, we assume that the capability distribution of strata, shifts at a bounded rate.

Assumption 3 (Population Drift)**.** The drift of stratum c is bounded by rate v_c for all t, t': $\sup_{x \in \mathcal{X}_L} |T_c(x; t) - T_c(x; t')| \leq v_c |t - t'|$, where $T_c(x; t)$ denotes $T_c(x)$ at time t.

Parameter $\beta \in (0, 1]$ functions as a sketch resolution. A stratum where the number of qualified agents for task τ at level x is below $n_c \beta$ is too sparse to be identified from the summary and is treated as invisible. Namely, a stratum is β-*visible* at $x \in \mathcal{X}_L$ if $T_c(x) \geq \beta$. When inspecting summaries, a smaller β allows the coordinator to detect capability even in sparse strata (costing a larger d), while a larger β restricts attention to strata with a higher proportion of capable agents, allowing more compact summaries (lower d), as Thm. 2 quantifies. We define the β-*visible portion* and the β-*visible top-q threshold* as:

$$T^{(\beta)}(x) := \sum_{c=1}^{K} \pi_c T_c(x) \mathbf{1}_{\{T_c(x) \geq \beta\}}, \quad x_q^{(\beta)} := \max\{x \in \mathcal{X}_L : T^{(\beta)}(x) \geq q\}. \tag{2}$$

Since $T^{(\beta)}(x)$ is non-increasing in x, the threshold $x_q^{(\beta)}$ is well-defined only when $T^{(\beta)}(0) \geq q$, and q sets the fraction of entities that the coordinator targets among entities in β-*visible* strata.

3.2 Sketch Construction

We instantiate summaries as additive sketches built from randomized vector representations called codewords (Kleyko et al., 2022; Rahimi & Recht, 2007; Kanerva, 2009). Each attribute (label) is mapped to a vector, and the sum of all member's attribute vectors forms the sketch of a stratum. Here, codewords are uniform draws from a torus. $\mathbb{T}^d := \{z \in \mathbb{C}^d : |z_j| = 1\,, \forall j\}$ is a torus, and z is uniform on $\mathbb{T}^d$ if its d coordinates are each drawn i.i.d. uniformly on the unit circle in $\mathbb{C}$.

Define the compression function $g : \mathcal{A} \to \Sigma$, where Σ is a finite set such that for every task $\tau \in \mathcal{T}$ there exist $F_\tau : \Sigma \to \{0, 1\}$ and $Q_\tau : \Sigma \to \mathcal{X}_L$, such that $f_\tau(a) = F_\tau(g(a))$ and $q_\tau(a) = Q_\tau(g(a))$, for all $a \in \mathcal{A}$. In other words, F_τ and Q_τ over Σ mirror the roles of f_τ and q_τ over $\mathcal{A}$. The *label* set Σ is sufficiently informative that the feasibility and quality of any entity can be recovered from its compressed attribute record $\sigma = g(a)$ alone.

Proposition 1. *Let $\{t_\sigma\}$ be a set of $|\Sigma|$ independently drawn random vectors from $\mathbb{T}^d$. The sketch of stratum c is created as $S_c := \sum_{i \in c} t_{g(a_{c,i})} \in \mathbb{C}^d$. For a task τ, with a threshold $x \in \mathcal{X}_L$ on quality, define the* ***qualified label set*** *$\Sigma_{\tau,x} := \{\sigma \in \Sigma : F_\tau(\sigma) = 1, Q_\tau(\sigma) \geq x\}$. The probe P_x^τ is designed specifically to detect this qualification: $P_x^\tau := \sum_{\sigma \in \Sigma_{\tau,x}} t_\sigma$. Utilizing this probe, an unbiased estimator of $T_c(x)$ can be obtained:*

$$\hat{T}_c(x) := \frac{Re\langle P_x^\tau, S_c \rangle}{d \cdot n_c}, \quad \langle u, v \rangle := \sum_{j=1}^{d} \bar{u}_j v_j \ ,$$

that is, $\mathbb{E}[\hat{T}_c(x)] = T_c(x)$, where the expectation is with respect to the choice of $\{t_\sigma\}$.

The proof, in App. A.3, expands $\langle P_x^\tau, S_c \rangle$ coordinate-wise: a term with $\sigma = g(a_{c,i})$ equals 1, and a term with $\sigma \neq g(a_{c,i})$ has zero mean because the two phases are independent and uniform. Moreover, the additivity of S_c allows for convenient updates when two strata join or separate or when an entity enters or leaves the stratum. For instance $S_{c' \cup c} = S_{c'} + S_c$.

The following lemma provides insight on T_c estimation error and the sketch dimension.

Lemma 1. *Define A as the maximum size among qualified label sets. $A := \max_{\tau,x} \left|\Sigma_{\tau,x}\right|$. Fix $\eta \in (0, 1]$ and suppose every sketch has dimension of $d \geq \frac{8A}{\eta^2} \log \frac{4K(L+1)}{\delta}$. For a task τ independent*

of the draw of $\{t_\sigma\}$, with probability of at least $1 - \frac{\delta}{2}$, the following holds for every stratum $c \in [K]$ and $x \in \mathcal{X}_L$:

$$\left|\hat{T}_c(x;t) - T_c(x;t')\right| \le \eta + v_c|t - t'| =: \eta_c,$$

with $t < t'$. Note that $\hat{T}_c$ uses the sketch built at time t, while T_c drifts as time passes.

The proof, in App. A.4, applies a Bernstein bound to the d i.i.d. coordinates of $\hat{T}_c(x)$, each of variance at most $A/2$ and bounded in $[-A, A]$, then a union bound over the $K(L+1)$ pairs (c, x), and the triangle inequality for the drift term.

3.3 Success Criteria: Coordinator Goal, Budget, and Sketch Dimension

This section defines when a coordinator's algorithm succeeds on a task, and what success costs in terms of individual queries and sketch dimension.

Definition 1 (Successful Discovery)**.** For a given target fraction $q \in (0, 1]$, sketch resolution $\beta \in (0, 1]$, failure probability $\delta \in (0, 1)$, and task $\tau \in \mathcal{T}$, a coordinator algorithm is (τ, q, β, δ)-*successful* if whenever $T^{(\beta)}(0) \ge q$, it outputs an entity $\hat{j}$ from some stratum c such that $f_\tau(a_{c,\hat{j}}) = 1$ and $q_\tau(a_{c,\hat{j}}) \ge x_q^{(\beta)}$ with probability at least $1 - \delta$.

That is, the returned entity must be able to perform the task and its performance quality reaches at least $x_q^{(\beta)}$, the highest quality level at which capable entities in β-*visible* strata still make up a q-fraction of the whole collective. This criterion, however, says nothing about cost. In the worst case, querying every entity is trivially (τ, q, β, δ)-*successful*, so cost is what separates algorithms. By Assum. 2, the coordinator has $2Kd$ real numbers of group-level summaries with no query cost, and computing *qualified portion* estimates from them requires only local arithmetic (Prop. 1). The following definition captures when a summary dimension suffices to keep queries within a budget.

Definition 2. For a task class $\mathcal{T}$, target fraction $q \in (0, 1]$, sketch resolution $\beta \in (0, 1]$, failure probability $\delta \in (0, 1)$, and query budget B, a sketch family of dimension d is *decision sufficient* if there exists an algorithm that is (τ, q, β, δ)-*successful* for all $\tau \in \mathcal{T}$ that uses at most B queries in expectation.

4 Sketch-Informed Elimination and Verification Algorithm

We propose the Sketch-Informed Elimination and Verification Algorithm (SIEVE, Alg. 1), which will be shown to be (τ, q, β, δ)-*successful*. A coordinator following SIEVE first fetches the summaries and derives the estimates $\hat{T}_c$ (Lines 1-2). Then comes the elimination step, which keeps only the strata that appear $\beta/2$-*visible* (Line 3). Lem. 1 justifies the choice of $\beta/2$ instead of β, since $\hat{T}_c(x)$ is used and not $T_c(x)$. For the same reason, the target fraction is relaxed from q to $3q/4$ and $\hat{x}$ plays the role of $x_q^{(\beta)}$ (Line 4-5). Line 6 then weights the surviving $\beta/2$-*visible* strata by their estimated qualifying mass $m_c := \pi_c \hat{T}_c(\hat{x})$, which is the estimated fraction of the whole collective that lies in stratum c and is qualified for the task at threshold $\hat{x}$. So queries concentrate where acceptable entities are more. Lines 7-11 repeatedly draw a stratum with probability proportional to its estimated qualifying mass and query it as in Assum. 1, testing whether the revealed entity meets the task requirements and the threshold. The algorithm outputs the entity if it meets both, and iterates otherwise. If no entity accepted after $Q_{\max} := \lceil \min\{32/(3q), 16/\beta\} \log(2/\delta) \rceil$ efforts, the algorithm declares failure (Line 12).

5 Theoretical Guarantees

We now show that SIEVE succeeds whenever every summary is accurate to within $\beta/4$, and bound the number of queries it uses.

Theorem 1 (SIEVE Guarantees)**.** *SIEVE(τ, q, β, δ) outputs an entity $\hat{j}$ with $f_\tau(a_{c,\hat{j}}) = 1$ and $q_\tau(a_{c,\hat{j}}) \ge x_q^{(\beta)}$ with probability at least $1 - \delta$, using at most $Q_{max} := \lceil \min\{\frac{32}{3q}, \frac{16}{\beta}\} \log \frac{2}{\delta} \rceil$ queries, if η_c (as in Lem. 1) is at most $\beta/4$ for all c and the problem is feasible (i.e. $T^{(\beta)}(0) \ge q$).*

Algorithm 1 SIEVE(τ, q, β, δ)

1: Read: fetch all strata summaries $S_1, \ldots, S_K$
2: Estimate: derive $\hat{T}_c(x)$ for every $x \in \mathcal{X}_L$ and $c \in [K]$. [Using S_c and Prop. 1]
3: **Eliminate**: for each $x \in \mathcal{X}_L$, set $D(x) := \{c : \hat{T}_c(x) \geq \frac{\beta}{2}\}$
4: Set $\hat{T}^{(\beta/2)}(x) := \sum_{c \in D(x)} \pi_c \hat{T}_c(x)$
5: Set $\hat{x} := \max\{x \in \mathcal{X}_L : \hat{T}^{(\beta/2)}(x) \geq 3q/4\}$, with $\hat{x} := 0$ if no such x exists.
6: Set $m_c := \pi_c \hat{T}_c(\hat{x})$ for $c \in D(\hat{x})$ and $m_c := 0$ otherwise.
7: **for** $t = 1$ **to** $Q_{\max}$ **do**
8: Sample stratum $c \sim \text{Categorical}\left(\frac{m_1}{\sum_k m_k}, \ldots, \frac{m_K}{\sum_k m_k}\right)$
9: **Verify**: Query c to receive a random un-queried entity's attribute $a_{c,\hat{j}}$. [Assum. 1]
10: **if** $f_\tau(a_{c,\hat{j}}) = 1$ and $q_\tau(a_{c,\hat{j}}) \geq \hat{x}$ **then**
11: **return** entity $\hat{j}$
12: Declare Failure

Proof. By Lem. 1, the condition $\eta_c \leq \beta/4$ holds when $\beta \geq 16\sqrt{\frac{2A}{d} \log \frac{4K(L+1)}{\delta}}$ and $\beta \geq 8\Delta_c v_c$ for all $c \in [K]$, where $\Delta_c := |t - t'|$ is the age of the sketch. Define $\beta_{\min}$ as the smallest β satisfying the condition. A coordinator with a target $\beta \geq \beta_{\min}$ may then run SIEVE. A smaller β generally needs a higher dimensional sketch $\uparrow d$ and/or a slower drift $\downarrow v_c$ to enable formal success guarantees. From Lem. 1 with probability at least $1-\delta/2$, every $c \in [K]$ and $x \in \mathcal{X}_L$ satisfies $\left|\hat{T}_c(x) - T_c(x)\right| \leq \eta_c \leq \beta/4$ (time parameter is removed for convenience). Condition on this event if $T_c(x) \geq \beta$, then $\hat{T}_c(x) \geq 3\beta/4 \geq \beta/2$. Also, if $\hat{T}_c(x) \geq \beta/2$, then $T_c(x) \geq \beta/4$. Hence, $D(x)$ (Alg. 1 Line 3) never discards a truly β-visible stratum and never retains a stratum with $T_c(x) < \beta/4$. Let $x^\star := x_q^{(\beta)}$. Plugging $x^\star$ into the definition of $D(x)$: every c with $T_c(x^\star) \geq \beta$ lies in $D(x^\star)$, so

$$\hat{T}^{(\beta/2)}(x^\star) = \sum_{c \in D(x^\star)} \pi_c \hat{T}_c(x^\star) \geq \sum_{c:\, T_c(x^\star) \geq \beta} \pi_c \big(T_c(x^\star) - \tfrac{\beta}{4}\big) \geq T^{(\beta)}(x^\star) - \tfrac{\beta}{4} \cdot \tfrac{T^{(\beta)}(x^\star)}{\beta} = \tfrac{3}{4} T^{(\beta)}(x^\star) \geq \tfrac{3q}{4},$$

using feasibility $T^{(\beta)}(x^\star) \geq q$. So $x^\star$ satisfies the defining inequality of $\hat{x} = \max\{x : \hat{T}^{(\beta/2)}(x) \geq 3q/4\}$, giving $\hat{x} \geq x^\star = x_q^{(\beta)}$. Since the algorithm accepts only entities verified against $f_\tau = 1$, $q_\tau \geq \hat{x}$ on the exact revealed record, every output satisfies $q_\tau(\hat{\jmath}) \geq \hat{x} \geq x_q^{(\beta)}$.

Write $T_{\text{in}} := \sum_{c \in D(\hat{x})} \pi_c T_c(\hat{x})$ and $\hat{m}_c := \pi_c \hat{T}_c(\hat{x})$, $\hat{M} := \sum_c \hat{m}_c$. For $c \in D(\hat{x})$: since $T_c(\hat{x}) \geq \beta/4$, $\hat{m}_c \leq \pi_c(T_c(\hat{x}) + \beta/4) \leq 2\pi_c T_c(\hat{x})$. For the lower comparison, if $\hat{T}_c(\hat{x}) \geq T_c(\hat{x})$, then $\hat{T}_c(\hat{x}) \geq T_c(\hat{x})/2$ trivially. Otherwise $T_c(\hat{x}) > \hat{T}_c(\hat{x}) \geq \beta/2$ by membership in $D(\hat{x})$, so $\beta/4 \leq T_c(\hat{x})/2$ and $\hat{T}_c(\hat{x}) \geq T_c(\hat{x}) - \beta/4 \geq T_c(\hat{x})/2$; either way $\hat{m}_c \geq \pi_c T_c(\hat{x})/2$. Summing, $\hat{M} \leq 2T_{\text{in}}$.

By Assumption 1, drawing stratum $c \in D(\hat{x})$ reveals a qualified member with probability at least $T_c(\hat{x})$, so the per-step success probability is

$$p = \sum_{c \in D(\hat{x})} \frac{\hat{m}_c}{\hat{M}} T_c(\hat{x}) \geq \frac{1}{2T_{\text{in}}} \sum_c \hat{m}_c T_c(\hat{x}) \geq \frac{1}{4T_{\text{in}}} \sum_c \pi_c T_c(\hat{x})^2.$$

Two bounds on this sum give the two terms in p: by $T_c(\hat{x}) \geq \beta/4$ on $D(\hat{x})$, $\sum_c \pi_c T_c(\hat{x})^2 \geq (\beta/4) T_{\text{in}}$, so $p \geq \beta/16$; by Cauchy-Schwarz with $\sum_{c \in D(\hat{x})} \pi_c \leq 1$, $\sum_c \pi_c T_c(\hat{x})^2 \geq T_{\text{in}}^2$, so $p \geq T_{\text{in}}/4$. Since $\hat{T}_c(\hat{x}) \leq 2T_c(\hat{x})$ on $D(\hat{x})$, we have $T_{\text{in}} \geq \hat{T}^{(\beta/2)}(\hat{x})/2 \geq 3q/8$, which gives $p \geq 3q/32$. Hence $p \geq \max\{3q/32,\ \beta/16\}$.

The number of draws until first success is stochastically dominated by a geometric random variable with parameter p, so

$$\mathbb{P}(\text{no success in } Q_{\max} \text{ draws}) \leq (1-p)^{Q_{\max}} \leq e^{-pQ_{\max}} \leq e^{-\log(2/\delta)} = \delta/2,$$

by the choice of $Q_{\max} \geq \log(2/\delta)/p$. One failure mode is Lem. 1 failing with probability at most $\delta/2$. Another failure could happen after $Q_{\max}$ unsuccessful draws which we showed happens with a probability at most $\delta/2$. Giving a total algorithm failure probability of at most δ. □

The proof above simultaneously shows *decision-sufficiency* (Def. 2) for any $d \geq \frac{8A}{\eta^2} \log \frac{4K(L+1)}{\delta}$ with $\eta \leq \beta/8$ and $8v_c\Delta_c \leq \beta$ (so that $\eta_c \leq \beta/4$ as required by Thm. 1), following SIEVE with budget $B = Q_{\max}$. It is worth noting that although $Q_{\max}$ bounds the maximum number of queries, the expected number of queries is smaller.

Remark 1 (Participation Rate). Define the *participation rate* at $\hat{x}$ as:

$$R(\hat{x}) := \frac{\sum_{c \in D(\hat{x})} \pi_c T_c(\hat{x})^2}{\sum_{c \in D(\hat{x})} \pi_c T_c(\hat{x})}$$

On the event of Lem. 1, the expected number of queries is at most $\min\{32/(3q), 4/R(\hat{x})\}$. If a single stratum c^* carries all qualifying mass (needle ensemble), i.e. $D(\hat{x}) = \{c^*\}$, then $R(\hat{x}) = T_{c^*}(\hat{x})$, which equals β when $T_{c^*}(\hat{x}) = \beta$, so the expected query count is at most $4/\beta$.

This remark shows when the summary is worth having. q lower bounds the chance of success per query under blind uniform sampling of entities. $R(\hat{x})$ is the average success rate over the retained strata after the summary has steered the search into them. If useful entities are concentrated, $R(\hat{x})$ can be much larger than q. In the worst case, however, order $1/q$ queries are inevitable in expectation even if every $T_c(x)$ is known, and any hard budget must be $\Omega(q^{-1} \log \delta^{-1})$ (App. A.8). Therefore, SIEVE's $Q_{\max}$ is optimal in q and δ up to constant factors. When concentration exists, summaries help more since every retained stratum has $T_c(\hat{x}) \geq \beta/4$, we have $4/R(\hat{x}) \leq 16/\beta$. In a needle ensemble (Def. 3), where a single stratum carries all qualifying mass at density β, this local cost is of order $1/\beta$ once the summary has located the stratum, whereas querying without a summary costs order K/β on average over which stratum carries it (App. A.9). The following theorem fixes the accuracy the summaries need and shows this dimension cannot be reduced.

Theorem 2 (decision-sufficient dimension)**.** *For fixed A, L, δ, and drift rates $8\Delta_c v_c \leq \beta$, the summary dimension needed to make the expected query cost of a (τ, q, β, δ)-successful coordinator independent of K is $d^* = \Theta(\beta^{-2} \log K)$. The upper bound holds for every population. For the lower bound, a needle ensemble (Def. 3) serves as a counterexample for every coordinator that reads each summary through a single noisy linear readout (Def. 4, App. A.10).*

Proof. For the upper bound, take $\eta = \beta/8$ in Lem. 1. Then $d \geq 512A\beta^{-2} \log \frac{4K(L+1)}{\delta}$. Under the conditions of the theorem and for fixed parameters, this $d = O(\beta^{-2} \log K)$. For the lower bound, see App. A.10. note that the needle ensemble concentrates qualified entities in a single stratum (unknown to coordinator). If $d \leq \alpha^2 \beta^{-2} \log K$ for some $\alpha \in (0, 1)$, then every successful procedure has an average expected query cost of order at least $K^{(1-\alpha)^2}/\beta$ over the hidden carrying stratum, which grows polynomially with K (App. A.10). This is the memory lower bound behind the method. A summary helps only if it can separate a true β signal from the largest accidental readout among many empty strata. With K empty strata, that accidental maximum is on the order of $\sqrt{\log K/d}$. Below the stated dimension, the carrying stratum is hidden among many plausible ones, so direct search again pays a large heterogeneity cost. □

6 EXPERIMENTS

This work makes three claims which we test in this section. First, a summary dimension of order $\beta^{-2} \log K$ is a threshold that the number of queries will not depend on the number of groups K and the collective size N, and below the threshold no coordinator that reads the summaries through a noisy linear readout can keep its cost from growing with K (Thm. 2). Second, above that threshold, the cost is set by how *concentrated* the capability is, about $1/R(\hat{x})$ queries, not by how *rare* it is, $1/q$ (Rem. 1). Third, whatever SIEVE returns is qualified with probability $1 - \delta$ (we used $\delta = 0.05$ throughout).

Test Populations. (1) *Synthetic needle* (Def. 3): tests $K = 16$ up to 512 strata of 2,000 entities each, so N ranges from 32,000 to 1.02 million. In one hidden stratum a fraction $\beta \in \{0.05, 0.1, 0.15, 0.2, 0.4\}$ of members is qualified and nothing else is, so the capability is rare overall ($q = \beta/K$) but dense where it lives. (2) *RapidAPI* (Qin et al., 2024): 16,464 web-API endpoints of the ToolBench crawl, grouped by their 49 published categories (e.g. Finance, Travel, Sports, eCommerce). An endpoint's record holds its category, pricing tier (free, freemium or paid), number of required input parameters, and its tool's popularity, which serves as quality ($L = 8$ quantile bins).

A task is a conjunction of requirements such as "a freemium Travel endpoint" or "a free endpoint with three required parameters", with a top-q target on popularity. We sample 24 tasks with q from $1.2 \cdot 10^{-4}$ to $8.4 \cdot 10^{-2}$ and use $\beta = 0.05$. (3) *Open LLM Leaderboard* (Fourrier et al., 2024): 4,496 evaluated language models, from which the coordinator chooses one (Shen et al., 2023). Models are grouped by publisher, with publishers of fewer than ten models pooled into one long-tail group so that a summary is more than a single directory entry ($K = 97$). A model's label combines license, size, type, chat template, and a mixture-of-experts flag (240 labels), and a task pairs a requirement with one of six benchmarks, such as "at most 8B parameters, chat or fine-tuned, top-q on GPQA". We sample 24 tasks with $q \in [10^{-3}, 0.2]$, $\beta = 0.1$.

Dimension Threshold. On the needle ensemble, blind search pays about K/β queries (Lem. 8) while a coordinator that knows the carrying group pays $1/\beta$, and a summary closes this gap only once the carrier's signal β exceeds the largest accidental readout among the K empty groups, about $\sqrt{\log K/d}$ (Rem. 3), that is, once d exceeds $d_{\mathrm{lb}} := \beta^{-2} \log K$, exactly where the lower bound of Lem. 9 stops applying. The experiment confirms that the threshold sits where the lower bound places it. Below d_{lb} the algorithm is relaxed to allow $40Q_{\max}$ queries, so the cost per success be visible. It uses 10^3 to $2 \cdot 10^4$ queries in that case. Above d_{lb}, curves fall onto the oracle's floor, with 9.3 to 10.3 queries against $1/\beta = 10$ for every K from 16 to 512, even as N grows 32-fold to a million entities (Fig. 1a). Under the hard budget $Q_{\max}$, success reaches 95% at 0.73 to $1.08\, d_{\mathrm{lb}}$ in nine (β, K) pairs (panel b), a constant close to one, even though the sufficient dimension of Thm. 2, which must hold for every population, is about 10^3 times larger. Panel (c) then moves from the needle towards a uniform population by spreading the same qualified entities over $k^\star$ groups: SIEVE's cost follows the prediction $1/R = k^\star/\beta$ and meets blind search only at $k^\star = K$, where the capability is everywhere and no summary can help. The mean cost settles somewhat later than success does, because it is dominated by rare cases in which an empty group survives elimination or the carrying group is eliminated.

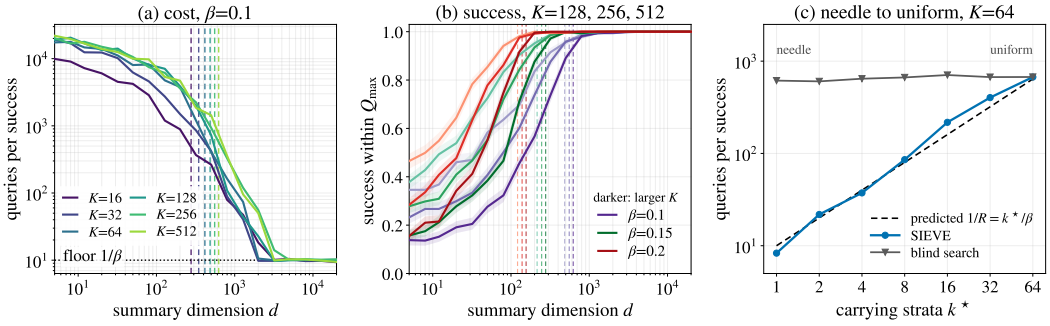


Figure 1: Once the summary dimension d passes $d_{\mathrm{lb}} = \beta^{-2} \log K$ (dashed lines), SIEVE's cost stops depending on the number of groups K. Below d_{lb}, noise in the summaries of the many empty groups hides the one group that holds the capability; above it, SIEVE finds a qualified entity in about $1/\beta$ queries (dotted line), as if it knew where to look. (a) Queries per success for $K = 16$ to 512 groups. (b) Chance of success within SIEVE's query budget. (c) As the capability spreads from one group to all K, SIEVE's cost follows its prediction $1/R$ and meets blind search only when the capability is everywhere.

Cost Prediction and Gain. Rem. 1 bounds SIEVE's expected cost by $4/R(\hat{x})$, and since R depends only on the registry and the task, a task's cost can be calculated before the search begins. Tab. 1 shows that this prediction is accurate on both registries, with SIEVE using close to $1/R$ queries on a typical task. A coordinator without summaries pays roughly $1/q$ instead, so the concentration R/q measures how much a summary can save, and SIEVE is 56 times cheaper than blind search on Open LLM Leaderboard. The saving comes from how the registry groups its entities rather than from the entities themselves. If RapidAPI's APIs are regrouped at random instead of by category, the capability is no longer concentrated in a few groups, SIEVE's cost rises from 8 to 59 queries, and its advantage over blind search falls from 170 to 23 times, while $1/R$ still predicts the cost.

Table 1: Predicting SIEVE's cost before searching and compare to blind search. Cost is the number of queries required to find a qualified entity on a typical task. Values denote the median over tasks, with brackets indicating the 95% confidence interval.

	Open LLM	RapidAPI (by category)	RapidAPI (at random)
queries predicted in advance ($1/R$)	4.4	5.8	86
queries SIEVE used	4.6 [3.5, 5.8]	8.0 [4.4, 11.6]	59 [50, 73]
queries blind search used	260 [103, 782]	1,360 [1140, 1690]	1,360 [1140, 1690]

SIEVE vs. Other Discovery Methods. We compare SIEVE with six alternatives, all working under the same assumptions (Assum. 1, 2, 3). Two of them use no summary. *Random search* samples entities from the whole collective, and *group scan* samples a few members of every group before committing to the most promising one, as in grouped quantile exploration (Lau et al., 2023). The other four read the same summaries as SIEVE but without its safety margins. *Greedy ranking* queries groups in order of their estimated share of qualified members, *weighted sampling* draws groups in proportion to that estimate, *Thompson sampling* treats the estimate as a prior and updates it after each miss, and *MF-LUCB* treats the summary as a cheap but biased observation of each group (Poiani et al., 2024).

Table 2: **Only SIEVE is both cheap and reliable.** Cost is the median over tasks of each method's queries divided by those of an oracle that knows every T_c and always queries the group with the highest remaining fraction of qualified members. A task is met when at least $1 - \delta = 95\%$ of its runs return a qualified entity, and "worst task" is the success rate on the hardest task.

	RapidAPI			Open LLM Leaderboard		
method	cost	tasks met	worst task	cost	tasks met	worst task
random search	163×	24/24	97%	194×	24/24	97%
group scan	486×	24/24	100%	1,510×	24/24	100%
greedy ranking	1.00×	13/24	47%	1.01×	9/24	23%
weighted sampling	1.63×	11/24	45%	3.88×	6/24	40%
Thompson sampling	1.00×	13/24	47%	1.01×	9/24	23%
MF-LUCB	1.00×	13/24	47%	1.01×	8/24	23%
SIEVE	1.13×	**24/24**	**100%**	2.39×	**24/24**	**100%**

Tab. 2 shows that SIEVE is the only method that is both cheap and reliable. It meets the target on every task of both registries at close to the oracle's cost, while the two methods without a summary are equally reliable but hundreds to thousands of times more expensive. The four methods that read the same summaries match the oracle's cost, but they meet the target on only half of the RapidAPI tasks and a third of the Open LLM Leaderboard tasks, and on their worst tasks most runs return an entity below the top-q benchmark. They lose reliability because they trust noisy estimates at face value, so a qualified group that looks slightly too sparse is dropped and the search settles for a weaker entity. SIEVE's margins prevent this, so its reliability is the guarantee of Thm. 1.

7 CONCLUSION

We studied capability discovery in large agent collectives, where a coordinator cannot afford to inspect every entity and instead relies on compact summaries that groups publish about their members. We showed that once these summaries are large enough to tell groups that contain the capability apart from those that do not, our algorithm SIEVE finds a qualified entity with a number of queries that no longer grows with the collective, and that below this size no coordinator reading the summaries can prevent such growth. The experiments further show that the benefit of a summary depends on how a registry groups its entities, that the cost of a search can be estimated from the registry before any query is made, and that the safety margins of SIEVE are what keep the returned entity above the quality benchmark when other methods reading the same summaries fall short. As open-ended agent networks scale, we hope this work paves the way toward principled, query-efficient discovery protocols for large decentralized collectives.

AI USE STATEMENT

In this work, we used generative AI tools for conceptualizing ideas, identifying related work, text editing, finding theoretical gaps, coding experiments, consistency checking, and organizing the Appendix section. All AI-assisted content was reviewed by the authors. We manually surveyed the literature to assess AI-suggested research ideas for potential overlap with prior work, tested AI-assisted code for correctness, and carried out proofs while consulting AI tools. We take responsibility for the final content of this work produced with the aid of generative AI.

REFERENCES

Pankaj K. Agarwal, Graham Cormode, Zengfeng Huang, Jeff M. Phillips, Zhewei Wei, and Ke Yi. Mergeable summaries. *ACM Trans. Database Syst.*, 38(4), 2013.

Matthew Aldridge, Oliver Johnson, and Jonathan Scarlett. Group testing: An information theory perspective. *Foundations and Trends in Communications and Information Theory*, 15(3-4):196–392, 2019.

Sepehr Assadi and Chen Wang. Exploration with limited memory: streaming algorithms for coin tossing, noisy comparisons, and multi-armed bandits. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, pp. 1237–1250. Association for Computing Machinery, 2020.

Sepehr Assadi and Chen Wang. The best arm evades: Near-optimal multi-pass streaming lower bounds for pure exploration in multi-armed bandits. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 311–358. PMLR, 2024.

Donald A Berry, Robert W Chen, Alan Zame, David C Heath, and Larry A Shepp. Bandit problems with infinitely many arms. *The Annals of Statistics*, pp. 2103–2116, 1997.

Burton H Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970.

Andrei Broder and Michael Mitzenmacher. Network applications of bloom filters: A survey. *Internet mathematics*, 1(4):485–509, 2004.

Alexandra Carpentier and Michal Valko. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning*, pp. 1133–1141. PMLR, 2015.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Weize Chen, Ziming You, Ran Li, Chen Qian, Chenyang Zhao, Cheng Yang, Ruobing Xie, Zhiyuan Liu, Maosong Sun, et al. Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence. In *International Conference on Learning Representations*, volume 2025, pp. 36374–36411, 2025.

Kenneth L Clarkson, Shashanka Ubaru, and Elizabeth Yang. Capacity analysis of vector symbolic architectures. *Journal of Artificial Intelligence Research*, 86, 2026.

Graham Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

Mohammadreza Doostmohammadian, Alireza Aghasi, Mohammad Pirani, Ehsan Nekouei, Houman Zarrabi, Reza Keypour, Apostolos I Rikos, and Karl H Johansson. Survey of distributed algorithms for resource allocation over multi-agent systems. *Annual Reviews in Control*, 59:100983, 2025.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *International Conference on Computational Learning Theory*, pp. 255–270. Springer, 2002.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open LLM leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`, 2024.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pp. 998–1027. PMLR, 2016.

Xiao-Yue Gong and Mark Sellke. Asymptotically optimal pure exploration for infinite-armed bandits. *arXiv preprint arXiv:2306.01995*, 2023.

Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, and Yann Traonmilin. Compressive statistical learning with random feature moments. *arXiv preprint arXiv:1706.07180*, pp. 1–38, 2017.

Shaolong Guo, Yuntao Wang, Zhou Su, Yanghe Pan, Qinnan Hu, and Tom H. Luan. Agent discovery in Internet of agents: Challenges and solutions. *IEEE Network*, pp. 1–9, 2026.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (MCP): Landscape, security threats, and future research directions. *ACM Transactions on Software Engineering and Methodology*, 35(10):1–37, 2026.

Yingwei Jin, Lulu Sun, Yanming Shen, Keqiu Li, and Geyong Min. A highly available spectrum allocation service model in dynamic spectrum market. *Future Generation Computer Systems*, 28 (6):940–946, 2012.

Kirthevasan Kandasamy, Gautam Dasarathy, Barnabas Poczos, and Jeff Schneider. The multi-fidelity multi-armed bandit. *Advances in neural information processing systems*, 29, 2016.

Pentti Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive computation*, 1(2):139–159, 2009.

Hideaki Kano, Junya Honda, Kentaro Sakamaki, Kentaro Matsuura, Atsuyoshi Nakamura, and Masashi Sugiyama. Good arm identification via bandit feedback. *Machine Learning*, 108(5): 721–745, 2019.

Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1781–1791. PMLR, 2020.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Nicolas Keriven, Anthony Bourrier, Rémi Gribonval, and Patrick Pérez. Sketching for large-scale learning of mixture models. *Information and Inference: A Journal of the IMA*, 7(3):447–508, 2018.

Denis Kleyko, Dmitri A Rachkovskij, Evgeny Osipov, and Abbas Rahimi. A survey on hyperdimensional computing aka vector symbolic architectures, part I: Models and data transformations. *ACM Computing Surveys*, 55(6):1–40, 2022.

Matthias Klusch, Patrick Kapahnke, Stefan Schulte, Freddy Lecue, and Abraham Bernstein. Semantic web service search: a brief survey. *KI-Künstliche Intelligenz*, 30(2):139–147, 2016.

Ivan Lau, Yan Hao Ling, Mayank Shrivastava, and Jonathan Scarlett. Max-quantile grouped infinite-arm bandits. In *International Conference on Algorithmic Learning Theory*, pp. 909–945. PMLR, 2023.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. A dynamic LLM-powered agent network for task-oriented agent collaboration. *arXiv preprint arXiv:2310.02170*, 2023.

Arnab Maiti, Vishakha Patil, and Arindam Khan. Multi-armed bandits with bounded arm-memory: Near-optimal guarantees for best-arm identification and regret minimization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19553–19565. Curran Associates, Inc., 2021.

Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

Elena Meshkova, Janne Riihijärvi, Marina Petrova, and Petri Mähönen. A survey on resource discovery mechanisms, peer-to-peer and service discovery frameworks. *Computer networks*, 52(11): 2097–2128, 2008.

Luca Muscariello, Vijoy Pandey, and Ramiz Polic. The agntcy agent directory service: Architecture and implementation. *arXiv preprint arXiv:2509.18787*, 2025.

Tony A Plate. Holographic reduced representations. *IEEE Transactions on Neural networks*, 6(3): 623–641, 1995.

Riccardo Poiani, Alberto Maria Metelli, and Marcello Restelli. Multi-fidelity best-arm identification. *Advances in Neural Information Processing Systems*, 35:17857–17870, 2022.

Riccardo Poiani, Rémy Degenne, Emilie Kaufmann, Alberto Maria Metelli, and Marcello Restelli. Optimal multi-fidelity best-arm identification. *Advances in Neural Information Processing Systems*, 37:121882–121927, 2024.

Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large language model-based multi-agent collaboration. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 41488–41505, 2025.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023.

Rao Surapaneni, Miku Jha, Michael Vakoc, and Todd Segal. Announcing the agent2agent protocol (a2a). *Google for Developers Blog*, 2025.

Ahmet Vedat Tokmak, Akhan Akbulut, and Cagatay Catal. Web service discovery: Rationale, challenges, and solution directions. *Computer Standards & Interfaces*, 88:103794, 2024.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. Multi-agent collaboration mechanisms: A survey of LLMs, 2025.

Jonas Wacker, Ruben Ohana, and Maurizio Filippone. Complex-to-real sketches for tensor products with applications to the polynomial kernel. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5181–5212. PMLR, 2023.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y Zou. Mixture-of-agents enhances large language model capabilities. In *International Conference on Learning Representations*, volume 2025, pp. 33944–33963, 2025.

Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. Multi-agent architecture search via agentic supernet. *arXiv preprint arXiv:2502.04180*, 2025a.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. Aflow: Automating agentic workflow generation. In *International Conference on Learning Representations*, volume 2025, pp. 34040–34077, 2025b.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. Language agents as optimizable graphs. *arXiv preprint arXiv:2402.16823*, 2024.

A Appendix

This appendix proves every theoretical claim of the main text. Tab. 3 lists where each one is proved. App. A.2–A.6 contain the upper bounds and App. A.7–A.11 the lower bounds. All lower bounds follow from one inequality, Lem. 4.

Table 3: Where each result of the main text is proved.

Main-text result	Proof
Rounding of real-valued scores (Sec. 3)	App. A.2
Unbiased estimator (Prop. 1)	App. A.3
Estimation error (Lem. 1)	App. A.4
SIEVE guarantees (Thm. 1)	App. A.5
Expected cost (Rem. 1) and decision sufficiency	App. A.6
Rarity floor and hard-budget floor (Sec. 5)	App. A.8
Search without summaries costs order K/β	App. A.9
Undersized summaries cost polynomially in K	App. A.10
Decision-sufficient dimension (Thm. 2)	App. A.11

A.1 Conventions

Randomness. There are three independent sources of randomness. The *codebook* $\{t_\sigma\}_{\sigma\in\Sigma}$ which is drawn once and the *query randomness* consists of the member orderings of Assum. 1 (each stratum is revealed in a uniformly random order, independently across strata) and any internal randomness U of the coordinator. The task τ and the population are fixed independently of the codebook. Probabilities are over both sources unless stated otherwise.

Time. The sketch of stratum c is built at time t and used at the query time t'. We write $\Delta_c := t' - t \geq 0$ for its age, so that $\eta_c = \eta + v_c\Delta_c$ in Lem. 1. The estimator $\hat{T}_c(x)$ is normalized by the size n_c at which S_c was built. True quantities (T_c, π_c, $T^{(\beta)}$, $x_q^{(\beta)}$) refer to time t', and the population does not change during a run. We drop time arguments when this causes no ambiguity.

Other conventions. log is the natural logarithm. We take $q, \beta \in (0, 1]$ and $\delta \in (0, 1)$. If $\sum_k m_k = 0$ after Line 6 of Alg. 1, SIEVE declares failure. Step 2 of App. A.5 shows this cannot happen on the event of Lem. 1.

A.2 Real-valued quality scores

Lemma 2 (Rounding)**.** *Let $\tilde{q}_\tau : \mathcal{A} \to [0, 1]$ be a real-valued quality and $q_\tau(a) := \lfloor L\tilde{q}_\tau(a)\rfloor/L \in \mathcal{X}_L$. Define $\tilde{T}_c(x)$ and $\tilde{T}^{(\beta)}(x)$ for $x \in [0, 1]$ as in equation 1 and equation 2 with $\tilde{q}_\tau$ in place of q_τ, and let $\tilde{x}_q := \sup\{x \in [0, 1] : \tilde{T}^{(\beta)}(x) \geq q\}$. Then:*

(i) $T_c(x) = \tilde{T}_c(x)$ for every c and every $x \in \mathcal{X}_L$. Hence $T^{(\beta)}$, feasibility, $x_q^{(\beta)}$, the sketch readouts and every run of SIEVE are the same under q_τ and $\tilde{q}_\tau$.

(ii) $\tilde{x}_q - \frac{1}{L} < x_q^{(\beta)} \leq \tilde{x}_q$.

(iii) Every entity with $f_\tau = 1$ and $q_\tau \geq x_q^{(\beta)}$, in particular every entity returned under Thm. 1, has $\tilde{q}_\tau > \tilde{x}_q - \frac{1}{L}$.

Proof. (i) Let $x = k/L$ with $k \in \{0, \dots, L\}$. Since k is an integer, $\lfloor L\tilde{q}_\tau(a) \rfloor \geq k$ if and only if $L\tilde{q}_\tau(a) \geq k$, so $q_\tau(a) \geq x \iff \tilde{q}_\tau(a) \geq x$. The counts in equation 1 therefore coincide on $\mathcal{X}_L$, and everything else in the claim depends on scores only through these counts. The label map g only needs to determine the rounded score.

(ii) The upper bound holds because $x_q^{(\beta)} \in [0,1]$ satisfies $\tilde{T}^{(\beta)}(x_q^{(\beta)}) = T^{(\beta)}(x_q^{(\beta)}) \geq q$. For the lower bound, we first show that $\{x : \tilde{T}^{(\beta)}(x) \geq q\}$ is the closed interval $[0, \tilde{x}_q]$. Each $\tilde{T}_c$ is non-increasing and left-continuous: if $x_k \uparrow x$, the finite sets $\{i : \tilde{q}_\tau(a_{c,i}) \geq x_k\}$ decrease to $\{i : \tilde{q}_\tau(a_{c,i}) \geq x\}$. Hence $\{x : \tilde{T}_c(x) \geq \beta\}$ is an interval $[0, x_c]$ closed on the right, and the indicator $\mathbf{1}\{\tilde{T}_c(x) \geq \beta\}$ is also non-increasing and left-continuous. Sums and products of non-negative, non-increasing, left-continuous functions keep both properties, so the superlevel set of $\tilde{T}^{(\beta)}$ at q is $[0, \tilde{x}_q]$. It is non-empty by feasibility. The grid point $\lfloor L\tilde{x}_q \rfloor / L$ lies in this interval, so $T^{(\beta)}(\lfloor L\tilde{x}_q \rfloor / L) \geq q$ by (i), and $x_q^{(\beta)} \geq \lfloor L\tilde{x}_q \rfloor / L > \tilde{x}_q - 1/L$.

(iii) $\tilde{q}_\tau \geq q_\tau \geq x_q^{(\beta)} > \tilde{x}_q - 1/L$. □

Rounding therefore costs less than $1/L$ in quality and nothing in queries. The dimension in Lem. 1 depends on L only through $\log(L+1)$ and through A.

A.3 Proof of Proposition 1

Write $(t_\sigma)_j = e^{i\phi_{\sigma,j}}$, where the phases $\phi_{\sigma,j}$ are i.i.d. uniform on $[0, 2\pi)$. Then

$$\langle P_x^\tau, S_c \rangle = \sum_{j=1}^{d} \sum_{\sigma \in \Sigma_{\tau,x}} \sum_{i \in c} e^{i(\phi_{g(a_{c,i}),j} - \phi_{\sigma,j})}.$$

A term with $g(a_{c,i}) = \sigma$ equals 1. A term with $g(a_{c,i}) \neq \sigma$ has mean $\mathbb{E}[e^{i\phi_{g(a_{c,i}),j}}]\,\mathbb{E}[e^{-i\phi_{\sigma,j}}] = 0$, because the two phases are independent and uniform. For each i, at most one $\sigma \in \Sigma_{\tau,x}$ equals $g(a_{c,i})$, and one does exactly when $f_\tau(a_{c,i}) = 1$ and $q_\tau(a_{c,i}) \geq x$. Hence $\mathbb{E}\langle P_x^\tau, S_c \rangle = d\, n_c T_c(x)$, which is real, and dividing the real part by $d\, n_c$ gives $\mathbb{E}[\hat{T}_c(x)] = T_c(x)$. □

A.4 Proof of Lemma 1

Fix c and x. Write $\Sigma^+ := \Sigma_{\tau,x}$, $n := n_c$, $m(\sigma) := |\{i \in c : g(a_{c,i}) = \sigma\}|$ and $T := T_c(x) = \frac{1}{n}\sum_{\sigma \in \Sigma^+} m(\sigma)$. If $\Sigma^+ = \emptyset$, then $\hat{T}_c(x) = T_c(x) = 0$, so let $|\Sigma^+| \geq 1$.

Coordinates. Fix a coordinate j and let ϕ_σ be the phase of $(t_\sigma)_j$. Then $(P_x^\tau)_j = \sum_{\sigma \in \Sigma^+} e^{i\phi_\sigma}$ and $(S_c)_j = \sum_{\sigma' \in \Sigma} m(\sigma') e^{i\phi_{\sigma'}}$, so $\hat{T}_c(x) = \frac{1}{d}\sum_{j=1}^{d} X_j$ with

$$X_j := \tfrac{1}{n}\mathrm{Re}\big(\overline{(P_x^\tau)_j}\,(S_c)_j\big) = T + \mathrm{Re}\, W_j, \qquad W_j := \frac{1}{n} \sum_{\sigma \in \Sigma^+} \sum_{\sigma' \neq \sigma} m(\sigma')\, e^{i(\phi_{\sigma'} - \phi_\sigma)},$$

since the diagonal terms $\sigma' = \sigma$ add up to T. The phases are independent across coordinates, so $X_1, \dots, X_d$ are i.i.d.

Mean and range. For $\sigma' \neq \sigma$ the phase difference is uniform, so $\mathbb{E}[W_j] = 0$ and $\mathbb{E}[X_j] = T$. Also $|W_j| \leq \frac{1}{n}\sum_{\sigma \in \Sigma^+}(n - m(\sigma)) = |\Sigma^+| - T \leq A$.

Variance. For independent uniform phases, $\mathbb{E}[e^{i\sum_\sigma k_\sigma \phi_\sigma}]$ is 1 if every integer k_σ is zero and 0 otherwise. Take index pairs (σ, σ') and (ρ, ρ') with $\sigma, \rho \in \Sigma^+$, $\sigma' \neq \sigma$ and $\rho' \neq \rho$. The phase $\phi_{\sigma'} - \phi_\sigma - \phi_{\rho'} + \phi_\rho$ cancels only if $(\rho, \rho') = (\sigma, \sigma')$, and $\phi_{\sigma'} - \phi_\sigma + \phi_{\rho'} - \phi_\rho$ cancels only if $(\rho, \rho') = (\sigma', \sigma)$, which requires $\sigma' \in \Sigma^+$. Hence

$$\mathbb{E}|W_j|^2 = \frac{1}{n^2} \sum_{\sigma \in \Sigma^+} \sum_{\sigma' \neq \sigma} m(\sigma')^2 \leq \sum_{\sigma \in \Sigma^+} \Big(1 - \frac{m(\sigma)}{n}\Big)^2 \leq |\Sigma^+|,$$

$$\mathbb{E}[W_j^2] = \frac{1}{n^2} \sum_{\substack{\sigma, \sigma' \in \Sigma^+ \\ \sigma \neq \sigma'}} m(\sigma) m(\sigma') \leq T^2.$$

Since $(\operatorname{Re} w)^2 = \frac{1}{2}|w|^2 + \frac{1}{2}\operatorname{Re}(w^2)$,

$$\operatorname{Var}(X_j) = \mathbb{E}[(\operatorname{Re} W_j)^2] \le \frac{|\Sigma^+| + T^2}{2} \le |\Sigma^+| \le A.$$

Concentration. Bernstein's inequality for the mean of d i.i.d. variables with $|X_j - T| \le b$ and $\operatorname{Var}(X_j) \le v$ gives

$$\mathbb{P}\big(|\hat{T}_c(x) - T_c(x)| > \eta\big) \le 2\exp\Big(-\frac{d\eta^2}{2v + \frac{2}{3}b\eta}\Big).$$

With $v, b \le A$ and $\eta \le 1$ the denominator is at most $\frac{8}{3}A$, so the bound is at most $2\exp(-d\eta^2/(4A))$. A union bound over the $K(L+1)$ pairs (c, x) gives total failure probability at most $2K(L+1)\exp(-d\eta^2/(4A))$, which is at most $\delta/2$ once $d \ge \frac{4A}{\eta^2}\log\frac{4K(L+1)}{\delta}$, and in particular under the condition of Lem. 1.

Drift. On this event $|\hat{T}_c(x;t) - T_c(x;t)| \le \eta$ for all (c, x). Assum. 3 gives $|T_c(x;t) - T_c(x;t')| \le v_c\Delta_c$, and the triangle inequality gives $|\hat{T}_c(x;t) - T_c(x;t')| \le \eta_c$. □

Remark 2 (Constants). The proof shows that $\frac{4A}{\eta^2}\log\frac{4K(L+1)}{\delta}$ already suffices, half the dimension stated in Lem. 1. A direct computation also gives $\operatorname{Var}(X_j) \le |\Sigma^+|/2 \le A/2$. We keep the simpler bound A, which changes only the constant.

A.5 Proof of Theorem 1

We prove the theorem under the hypothesis

(H) $d \ge \frac{8A}{\eta^2}\log\frac{4K(L+1)}{\delta}$ for some $\eta \in (0, 1]$, $\eta_c = \eta + v_c\Delta_c \le \beta/4$ for every c, and $T^{(\beta)}(0) \ge q$.

A sufficient condition is $\eta = \beta/8$, i.e. $d \ge 512A\beta^{-2}\log\frac{4K(L+1)}{\delta}$, together with $8v_c\Delta_c \le \beta$ for all c. Let $\mathcal{E}$ be the event of Lem. 1, on which $|\hat{T}_c(x) - T_c(x)| \le \beta/4$ for all (c, x). It depends only on the codebook, and $\mathbb{P}(\mathcal{E}^c) \le \delta/2$. Until Step 7 we fix a codebook in $\mathcal{E}$. Then $D(\cdot)$, $\hat{x}$ and the weights m_c are fixed, and only the query randomness remains.

Step 1 (elimination). Fix $x \in \mathcal{X}_L$. If $T_c(x) \ge \beta$, then $\hat{T}_c(x) \ge 3\beta/4 \ge \beta/2$, so $c \in D(x)$. If $c \in D(x)$, then $T_c(x) \ge \hat{T}_c(x) - \beta/4 \ge \beta/4$.

Step 2 ($\hat{x} \ge x^\star$). Let $x^\star := x_q^{(\beta)}$, which exists by feasibility. By Step 1, $D(x^\star)$ contains every stratum with $T_c(x^\star) \ge \beta$. The other strata in $D(x^\star)$ have $\hat{T}_c(x^\star) \ge \beta/2 > 0$, so dropping them only decreases the sum:

$$\begin{aligned}\hat{T}^{(\beta/2)}(x^\star) &\ge \sum_{c:\, T_c(x^\star)\ge\beta} \pi_c\big(T_c(x^\star) - \tfrac{\beta}{4}\big) = T^{(\beta)}(x^\star) - \tfrac{\beta}{4}\sum_{c:\, T_c(x^\star)\ge\beta}\pi_c \\ &\ge T^{(\beta)}(x^\star) - \tfrac{\beta}{4}\cdot\frac{T^{(\beta)}(x^\star)}{\beta} = \tfrac{3}{4}T^{(\beta)}(x^\star) \ge \tfrac{3q}{4},\end{aligned}$$

where the second inequality uses $\pi_c \le \pi_c T_c(x^\star)/\beta$ for these strata. So $x^\star$ is in the set that defines $\hat{x}$. Hence $\hat{x} \ge x^\star$, the fallback $\hat{x} := 0$ is never used, and

$$\hat{M} := \sum_c m_c = \hat{T}^{(\beta/2)}(\hat{x}) \ge \tfrac{3q}{4} > 0. \tag{3}$$

In particular $D(\hat{x}) \ne \emptyset$, and the distribution in Line 8 is well defined.

Step 3 (every output is correct). SIEVE returns only an entity whose revealed record satisfies $f_\tau = 1$ and $q_\tau \ge \hat{x} \ge x^\star$, which is the success criterion of Def. 1. So on $\mathcal{E}$ the only way to fail is to declare failure after $Q_{\max}$ misses.

Step 4 (weights). Write $T_c := T_c(\hat{x})$ and $T_{\text{in}} := \sum_{c\in D(\hat{x})}\pi_c T_c$. For $c \in D(\hat{x})$, Step 1 gives $T_c \ge \beta/4$, so $\hat{T}_c(\hat{x}) \le T_c + \beta/4 \le 2T_c$. For the lower comparison, if $\hat{T}_c(\hat{x}) \ge T_c$ then $\hat{T}_c(\hat{x}) \ge T_c/2$

trivially. Otherwise $T_c > \hat{T}_c(\hat{x}) \geq \beta/2$, so $\beta/4 < T_c/2$ and $\hat{T}_c(\hat{x}) \geq T_c - \beta/4 \geq T_c/2$. In both cases

$$\tfrac{1}{2}\pi_c T_c \leq m_c \leq 2\pi_c T_c \quad (c \in D(\hat{x})), \qquad \text{hence} \qquad \tfrac{1}{2}\hat{M} \leq T_{\text{in}} \leq 2\hat{M}. \tag{4}$$

Step 5 (success probability of each draw). Let H_k be the history after k misses. At draw $k+1$, stratum c is chosen with probability $m_c/\hat{M}$, independently of H_k. By Assum. 1, the revealed record is uniform over the u_c members of c not yet revealed. Every member revealed so far failed the acceptance test $\{f_\tau = 1,\, q_\tau \geq \hat{x}\}$, which is exactly the event counted by T_c, so all $n_c T_c$ qualified members of c are still unrevealed. For $c \in D(\hat{x})$ this is a positive integer, since $T_c \geq \beta/4 > 0$. Hence $u_c \geq n_c T_c \geq 1$, the stratum is never exhausted before a success, and the revealed member is qualified with probability $n_c T_c/u_c \geq T_c$. Therefore

$$\begin{aligned} \mathbb{P}(\text{success at draw } k+1 \mid H_k) \;&\geq\; p := \sum_{c \in D(\hat{x})} \frac{m_c}{\hat{M}}\, T_c \\ &\geq\; \frac{1}{2T_{\text{in}}} \sum_{c \in D(\hat{x})} m_c T_c \;\geq\; \frac{1}{4T_{\text{in}}} \sum_{c \in D(\hat{x})} \pi_c T_c^2 = \frac{R(\hat{x})}{4}, \end{aligned} \tag{5}$$

using $\hat{M} \leq 2T_{\text{in}}$ and $m_c \geq \frac{1}{2}\pi_c T_c$ from equation 4.

Step 6 (two lower bounds on p). Since $T_c \geq \beta/4$ on $D(\hat{x})$, $\sum_c \pi_c T_c^2 \geq \frac{\beta}{4} T_{\text{in}}$, so $R(\hat{x}) \geq \beta/4$ and $p \geq \beta/16$. By Cauchy–Schwarz with $\sum_{c \in D(\hat{x})} \pi_c \leq 1$, $T_{\text{in}}^2 \leq \sum_c \pi_c T_c^2$, so $R(\hat{x}) \geq T_{\text{in}} \geq \hat{M}/2 \geq 3q/8$ by equation 4 and equation 3, and $p \geq 3q/32$. Together,

$$p \;\geq\; \max\Big\{\frac{3q}{32},\, \frac{R(\hat{x})}{4}\Big\} \;\geq\; \max\Big\{\frac{3q}{32},\, \frac{\beta}{16}\Big\} = \frac{1}{\min\{32/(3q),\, 16/\beta\}}. \tag{6}$$

Step 7 (failure probability). By the chain rule and equation 5,

$$\mathbb{P}(Q_{\max} \text{ misses} \mid \text{codebook}) = \prod_{k=0}^{Q_{\max}-1} \mathbb{P}(\text{miss at } k+1 \mid H_k) \leq (1-p)^{Q_{\max}} \leq e^{-pQ_{\max}} \leq \frac{\delta}{2},$$

because $Q_{\max} \geq \min\{32/(3q), 16/\beta\} \log(2/\delta)$ and equation 6 give $pQ_{\max} \geq \log(2/\delta)$. Removing the conditioning and using Step 3,

$$\mathbb{P}(\text{SIEVE fails}) \leq \mathbb{P}(\mathcal{E}^c) + \mathbb{E}\big[\mathbf{1}_{\mathcal{E}}\, \mathbb{P}(Q_{\max} \text{ misses} \mid \text{codebook})\big] \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

By construction, SIEVE never makes more than $Q_{\max}$ queries. □

A.6 Proof of Remark 1 and decision sufficiency

Lemma 3 (Expected cost)**.** *Under (H), for every codebook in $\mathcal{E}$,*

$$\mathbb{E}[Q \mid \textit{codebook}] \leq \min\Big\{Q_{\max},\, \frac{32}{3q},\, \frac{4}{R(\hat{x})}\Big\}.$$

Unconditionally, $\mathbb{E}[Q] \leq Q_{\max} \leq \lceil 16\beta^{-1} \log(2/\delta) \rceil$, which depends on neither K nor N.

Proof. By equation 5, $\mathbb{P}(Q > k \mid \text{codebook}) \leq (1-p)^k$, so $\mathbb{E}[Q \mid \text{codebook}] = \sum_{k=0}^{Q_{\max}-1} \mathbb{P}(Q > k \mid \text{codebook}) \leq \min\{Q_{\max}, 1/p\}$, and equation 6 bounds $1/p$. The unconditional bound holds because $Q \leq Q_{\max}$ on every run. □

This is the expected-cost claim of Rem. 1. For the needle claim: if a single stratum c^* carries all qualifying mass, every other stratum has $T_c(\hat{x}) = 0 < \beta/4$ and is removed by Step 1, so $D(\hat{x}) = \{c^*\}$ on $\mathcal{E}$ and $R(\hat{x}) = T_{c^*}(\hat{x})$. The discussion after Rem. 1 uses two more facts. Step 6 shows $R(\hat{x}) \geq \beta/4$, so $4/R(\hat{x}) \leq 16/\beta$. A query to a uniformly random member of the whole collective succeeds with probability $T(x^\star) \geq T^{(\beta)}(x^\star) \geq q$, which is the blind success rate that q lower-bounds.

Proposition 2 (Decision sufficiency). *Let every task in $\mathcal{T}$ be fixed independently of the codebook. If*

$$d \;\geq\; \frac{512A}{\beta^2}\log\frac{4K(L+1)}{\delta} \qquad \textit{and} \qquad 8v_c\Delta_c \leq \beta \;\textit{for all } c,$$

then SIEVE is (τ, q, β, δ)-successful for every $\tau \in \mathcal{T}$ and uses at most $Q_{\max}$ queries on every run. Hence the sketch family of dimension d is decision sufficient (Def. 2) with budget $B = Q_{\max}$.

Proof. Take $\eta = \beta/8 \leq 1$. The dimension condition is then that of Lem. 1, and $\eta_c = \beta/8 + v_c\Delta_c \leq \beta/4$, so (H) holds for every feasible τ and Thm. 1 applies to each τ. Def. 2 asks for success on each task separately, so no union bound over $\mathcal{T}$ is needed. □

A.7 Lower bounds: setting and two tools

Procedures. A *procedure* has side information I (the registry, plus whatever summaries or readouts the model allows) and internal randomness U, both independent of the member orderings of Assum. 1. At each step it either queries a stratum or stops, based on I, U and the records revealed so far. When it stops, it outputs a revealed entity, a *blind pick* from some stratum, or nothing. Records are assigned to members exchangeably, so an unrevealed member carries no information, and a blind pick from stratum c is without loss of generality a uniformly random unrevealed member of c. Q denotes the number of queries. SIEVE outputs only revealed entities; we allow blind picks so that the lower bounds are as strong as possible.

Background instances. An instance is a *background instance* if each stratum c has a fraction $\rho_c \in [0, 1]$ of *qualified* members (members that meet the success criterion of Def. 1) and all its other members share one infeasible record b_c. The *null run* is the run of the procedure, with the actual (I, U), in which every query to c returns b_c. For an integer $s \geq 0$, define on the null run:

- $N_c(s)$, the number of queries to c among the first s queries;
- $G_s \in [K] \cup \{\emptyset\}$, the stratum of the blind pick if the null run stops within s queries with a blind pick, and $\emptyset$ otherwise;
- the *load* $\ell_c(s) := N_c(s) + 2 \cdot \mathbf{1}\{G_s = c\}$.

These depend only on (I, U), and $\sum_c \ell_c(s) \leq s + 2$.

Definition 3 (Needle ensemble). Fix $K \geq 2$, $\beta \in (0, 1]$ and a common stratum size n with $\beta n \in \mathbb{N}$, so that $\pi_c = 1/K$. Let $a^\star$ be a record with $f_\tau(a^\star) = 1$ and $q_\tau(a^\star) = 1$, and let $b_1, \ldots, b_K$ be records with $f_\tau(b_c) = 0$ whose labels $g(b_1), \ldots, g(b_K)$ are distinct from each other and from $g(a^\star)$. In instance $E_{c^\star}$, the carrying stratum $c^\star$ has βn members with record $a^\star$ and $(1-\beta)n$ with record $b_{c^\star}$, and every other stratum c has n members with record b_c.

In $E_{c^\star}$, $T_{c^\star}(x) = \beta$ and $T_c(x) = 0$ for $c \neq c^\star$ at every $x \in \mathcal{X}_L$, so $T^{(\beta)}(x) = \beta/K$. We take $q = \beta/K$. Then the instance is feasible, $x_q^{(\beta)} = 1$, and success means returning a member with record $a^\star$. Each $E_{c^\star}$ is a background instance with $\rho_c = \beta\,\mathbf{1}\{c = c^\star\}$, and $\Sigma_{\tau,x} = \{g(a^\star)\}$ for every x, so $A = 1$. The labels of the backgrounds are distinct so that readout errors are independent across strata (Rem. 3); with a shared background, all non-carrying strata would have the same sketch.

Lemma 4 (Null-run inequality). *Consider a background instance and an integer $s \geq 0$ with $n_c \geq 2s$ for every c. Every procedure satisfies*

$$\mathbb{P}(\textit{success}) \;\leq\; \mathbb{P}(Q > s) + \mathbb{E}\Big[\sum_c \rho_c\,\ell_c(s)\Big] \;\leq\; \mathbb{P}(Q > s) + (s+2)\max_c \rho_c.$$

Proof. Condition on (I, U). The real run coincides with the null run until the first query that reveals a qualified member. Let F be the event that a qualified member is revealed within the first s queries. Then F is the event that, for some c, one of the first $N_c(s)$ positions in the ordering of c holds a qualified member. The ordering is uniform and independent of (I, U), so each position holds a qualified member with probability ρ_c, and a union bound gives $\mathbb{P}(F \mid I, U) \leq \sum_c \rho_c N_c(s)$.

On $F^c \cap \{Q \leq s\}$ the real run is the null run until it stops, so every revealed entity is unqualified, and success requires a qualified blind pick from $G_s \neq \emptyset$. Given F^c, the $N_{G_s}(s) \leq s \leq n_{G_s}/2$

revealed members of G_s are unqualified, so all $\rho_{G_s} n_{G_s}$ qualified members are among the at least $n_{G_s}/2$ unrevealed ones, and the pick is qualified with probability at most $2\rho_{G_s}$.

Hence $\mathbb{P}(\text{success}, Q \le s \mid I, U) \le \sum_c \rho_c N_c(s) + 2\rho_{G_s} = \sum_c \rho_c \ell_c(s)$. Taking expectations and adding $\mathbb{P}(Q > s)$ gives the first inequality, and $\sum_c \ell_c(s) \le s + 2$ gives the second. □

Lemma 5 (Tail sum). *Let $Q \ge 0$ be integer-valued and $\kappa, h > 0$. If $\mathbb{P}(Q > s) \ge \kappa - hs$ for every integer $0 \le s \le \kappa/h$, then $\mathbb{E}[Q] \ge \kappa^2/(2h)$.*

Proof. Let $S := \lfloor \kappa/h \rfloor$. Then $\mathbb{E}[Q] = \sum_{s \ge 0} \mathbb{P}(Q > s) \ge \sum_{s=0}^{S} (\kappa - hs) = (S+1)(\kappa - hS/2)$. Since $S + 1 > \kappa/h$ and $hS \le \kappa$, this is at least $\frac{\kappa}{h} \cdot \frac{\kappa}{2}$. □

A.8 Rarity floor and hard-budget floor

Lemma 6 (Rarity floor). *Let $2q < 1 - \delta$. There is a single-stratum instance with qualified fraction q on which every (τ, q, β, δ)-successful procedure, even one that knows all tail functions exactly, has $\mathbb{E}[Q] \ge (1 - \delta - 2q)^2/(2q) = \Omega(q^{-1})$.*

Proof. Take $K = 1$, $\beta \le q$, and $n \ge 2/q$ members with $qn \in \mathbb{N}$. Of these, qn have a record $a^\star$ with $f_\tau(a^\star) = 1$ and $q_\tau(a^\star) = 1$, and the rest have an infeasible record b. Then $T_1(x) = q \ge \beta$ for every x, the instance is feasible, $x_q^{(\beta)} = 1$, and success means returning a member with record $a^\star$. The tail functions are known, and any summary is a function of the fixed multiset of records, so it adds no information. This is a background instance with $\rho_1 = q$. For every integer $0 \le s \le n/2$, Lem. 4 and success give $1 - \delta \le \mathbb{P}(Q > s) + q(s+2)$, i.e., $\mathbb{P}(Q > s) \ge (1 - \delta - 2q) - qs$. Lem. 5 with $\kappa = 1 - \delta - 2q$ and $h = q$ gives the claim; the range $s \le \kappa/h \le 1/q$ lies within $s \le n/2$. □

On this instance, querying until success costs $O(1/q)$ in expectation, so the expected cost need not grow with $\log(1/\delta)$. The $\log(1/\delta)$ factor in $Q_{\max}$ is the price of a hard budget, i.e., a bound on every run:

Lemma 7 (Hard-budget floor). *Let $q \in (0, 1/2)$ and take the instance of Lem. 6 with $n \ge \log(1/\delta)/q$. Every (τ, q, β, δ)-successful procedure that makes at most B queries on every run satisfies*

$$B \;\ge\; \frac{(1 - 2q)\log(1/\delta)}{2q} - 1.$$

Proof. If $B > n/2$, then $B > \log(1/\delta)/(2q)$ and the claim holds, so assume $B \le n/2$. Let F_0 be the event that none of the first B positions of the ordering holds a qualified member. Then

$$\mathbb{P}(F_0) = \prod_{k=0}^{B-1} \left(1 - \frac{qn}{n-k}\right) \ge \left(1 - \frac{qn}{n-B}\right)^B \ge (1 - 2q)^B.$$

On F_0 every query reveals b, so a revealed output fails, and a blind pick is uniform over at least $n/2$ unrevealed members and is qualified with probability at most $2q$. Hence $\delta \ge \mathbb{P}(\text{failure}) \ge \mathbb{P}(F_0)(1 - 2q) \ge (1 - 2q)^{B+1}$. Taking logarithms and using $\log \frac{1}{1-y} \le \frac{y}{1-y}$ for $y \in [0, 1)$ gives $B + 1 \ge \frac{\log(1/\delta)}{\log(1/(1-2q))} \ge \frac{(1-2q)\log(1/\delta)}{2q}$. □

With Thm. 1, this shows that SIEVE's budget $Q_{\max} = O(q^{-1} \log \delta^{-1})$ is optimal in q and δ, up to constant factors, among procedures with a hard budget.

A.9 Search without summaries

Lemma 8 (Cost without summaries). *On the needle ensemble with $q = \beta/K$ and $n \ge 2K/\beta$, let a procedure use only the registry and queries, and let it be (τ, q, β, δ)-successful on every $E_{c^\star}$. Then*

$$\frac{1}{K} \sum_{c^\star = 1}^{K} \mathbb{E}_{c^\star}[Q] \;\ge\; \frac{(1 - \delta - 2\beta/K)^2\, K}{2\beta}.$$

Proof. All strata have size n, so the registry is the same under every $E_{c^\star}$, and the loads $\ell_c(s)$ depend on U alone, with the same law under every instance. For each $c^\star$ and every integer $s \le n/2$, Lem. 4 with $\rho_c = \beta \mathbf{1}\{c = c^\star\}$ gives $1 - \delta \le \mathbb{P}_{c^\star}(Q > s) + \beta\, \mathbb{E}[\ell_{c^\star}(s)]$. Averaging over $c^\star$,

$$\frac{1}{K} \sum_{c^\star} \mathbb{P}_{c^\star}(Q > s) \;\ge\; 1 - \delta - \frac{\beta}{K}\, \mathbb{E}\Big[\sum_{c} \ell_c(s)\Big] \;\ge\; \Big(1 - \delta - \frac{2\beta}{K}\Big) - \frac{\beta}{K}\, s.$$

The left-hand side is $\mathbb{P}(Q > s)$ when $c^\star$ is uniform. Lem. 5 with $\kappa = 1 - \delta - 2\beta/K$ and $h = \beta/K$ gives the claim; the range $s \le K/\beta$ lies within $s \le n/2$. □

The bound is tight in order: querying a uniformly random stratum succeeds with probability at least β/K at every draw (sampling is without replacement), so it needs at most K/β queries in expectation. Without summaries the cost is therefore of order K/β, whereas SIEVE pays at most $4/\beta$ once the summary has located the carrier (Lem. 3). The worst case over $c^\star$ is at least the average.

A.10 Undersized summaries bring back the K cost

A lower bound on the sketch dimension needs a model of how summaries are read. Rem. 4 shows that an unrestricted decoder of an exact torus sketch can find the carrier at $d = 1$. We therefore use the model that covers SIEVE and every probe-based coordinator: one noisy linear readout per stratum.

Definition 4 (Readout-noise model)**.** On the needle ensemble, the side information is the registry and one readout per stratum,

$$Y_c = T_c(1) + \sigma_d\, Z_c, \qquad c \in [K],$$

where $Z_1, \dots, Z_K$ are i.i.d. standard Gaussians, independent of U and of the orderings, and $\sigma_d > 0$ is the readout noise of a dimension-d sketch. Since $T_c(x)$ does not depend on x in the ensemble, the same Y_c serves at every x.

Lemma 9 (Undersized summaries)**.** *In the readout-noise model with $n \ge 2K/\beta$, suppose*

$$\frac{\beta}{\sigma_d} \;\le\; \alpha\sqrt{2\log K} \qquad \textit{for some } \alpha \in (0,1). \tag{7}$$

Every procedure that is $(\tau, \beta/K, \beta, \delta)$-successful on every $E_{c^\star}$ satisfies

$$\frac{1}{K} \sum_{c^\star=1}^{K} \mathbb{E}_{c^\star}[Q] \;\ge\; \frac{\big(1 - \delta - 3\beta K^{-(1-\alpha)^2}\big)^2}{3\beta}\, K^{(1-\alpha)^2}.$$

In particular, if $K^{(1-\alpha)^2} \ge 6\beta/(1-\delta)$, the average is at least $\frac{(1-\delta)^2}{12\beta} K^{(1-\alpha)^2}$.

Proof. Let $y_c := Y_c/\sigma_d$ and $\mu := \beta/\sigma_d \le \alpha\sqrt{2\log K}$. Under E_c, $y = \mu\, e_c + Z$. Let $\mathbb{P}_0$ be the law of $y = Z$ (no carrier), used only as a reference measure. The likelihood ratio of E_c to $\mathbb{P}_0$ is $\Lambda_c(y) = \exp(\mu y_c - \mu^2/2)$.

The registry is fixed, so each load is a function $\ell_c(s) = \ell_c(s; y, U)$. For each c and every $s \le n/2$, Lem. 4 gives $1 - \delta \le \mathbb{P}_c(Q > s) + \beta\, \mathbb{E}_c[\ell_c(s)]$. Change measure and split on $\{\Lambda_c \le M\}$ for a threshold $M > 0$. Using $\ell_c(s) \le s + 2$ and $\mathbb{E}_0[\Lambda_c \mathbf{1}\{\Lambda_c > M\}] = \mathbb{P}_c(\Lambda_c > M)$,

$$\mathbb{E}_c[\ell_c(s)] = \mathbb{E}_0[\Lambda_c\, \ell_c(s)] \le M\, \mathbb{E}_0[\ell_c(s)] + (s+2)\, \mathbb{P}_c(\Lambda_c > M).$$

By symmetry $\mathbb{P}_c(\Lambda_c > M)$ does not depend on c. Averaging over c and using $\sum_c \ell_c(s) \le s + 2$,

$$\frac{1}{K} \sum_{c} \mathbb{E}_c[\ell_c(s)] \;\le\; (s+2)\Big(\frac{M}{K} + \mathbb{P}_1(\Lambda_1 > M)\Big). \tag{8}$$

Choose $M := K^\gamma$ with $\gamma := 1 - (1-\alpha)^2 = \alpha(2-\alpha) \in (0,1)$, so $M/K = K^{-(1-\alpha)^2}$. Under E_1, $y_1 = \mu + Z_1$, so $\Lambda_1 > M$ if and only if $Z_1 > z := \gamma \log K/\mu - \mu/2$. Since z decreases in μ, equation 7 gives

$$z \;\ge\; \frac{\gamma \log K}{\alpha\sqrt{2\log K}} - \frac{\alpha\sqrt{2\log K}}{2} = (1-\alpha)\sqrt{2\log K} > 0,$$

and the Gaussian tail bound $\mathbb{P}(Z > z) \le \frac{1}{2}e^{-z^2/2}$ gives $\mathbb{P}_1(\Lambda_1 > M) \le \frac{1}{2}K^{-(1-\alpha)^2}$. Substituting into equation 8 and averaging the null-run inequality over $c^\star$,

$$\frac{1}{K}\sum_{c^\star}\mathbb{P}_{c^\star}(Q > s) \;\ge\; (1-\delta-2h) - hs, \qquad h := \tfrac{3}{2}\beta K^{-(1-\alpha)^2}.$$

Lem. 5 with $\kappa = 1-\delta-2h$ gives the first bound; the range $s \le \kappa/h \le \frac{2}{3}K^{(1-\alpha)^2}/\beta$ lies within $s \le n/2$ because $n \ge 2K/\beta$. If $3\beta K^{-(1-\alpha)^2} \le (1-\delta)/2$, the numerator is at least $(1-\delta)^2/4$, which gives the second bound. □

The exponent $(1-\alpha)^2$ is positive for every $\alpha < 1$, so the cost grows polynomially in K. As $\alpha \to 0$, the bound recovers the order K/β of Lem. 8.

Remark 3 (Calibration to the torus sketch). On the needle ensemble, $A = 1$ and $P_x^\tau = t_{g(a^\star)}$. The readout of a non-carrying stratum is $\hat{T}_c = \frac{1}{d}\sum_{j=1}^{d}\cos\psi_{c,j}$ with $\psi_{c,j} := \phi_{g(b_c),j} - \phi_{g(a^\star),j}$. Given $t_{g(a^\star)}$, the $\psi_{c,j}$ are i.i.d. uniform across c and j because the background labels are distinct. So these readouts are independent across strata, with mean 0 and variance $1/(2d)$, and by the central limit theorem close to $\mathcal{N}(0, 1/(2d))$. The carrier's readout is β plus $(1-\beta)$ times an independent copy, so its noise is smaller by the factor $1-\beta$; the model uses the same σ_d for every stratum. The calibrated noise level is $\sigma_d = (2d)^{-1/2}$, and equation 7 becomes

$$d \;\le\; \alpha^2\beta^{-2}\log K,$$

the form used in Thm. 2. At the coarser scale $\sigma_d = d^{-1/2}$ it becomes $d \le 2\alpha^2\beta^{-2}\log K$; either way the threshold is of order $\beta^{-2}\log K$. The largest accidental readout among K empty strata is about $\sigma_d\sqrt{2\log K} = \sqrt{\log K/d}$. With a shared background, all non-carrier readouts would be one repeated draw, and the $\log K$ would disappear.

Remark 4 (Scope of the lower bound). The restriction to one noisy readout per stratum cannot be dropped. With the public codebook, a coordinator that decodes an exact torus sketch recognizes the carrier at any d: for $\beta < 1$, $|(S_c)_1| = n$ for every non-carrier, while $|(S_{c^\star})_1| = |\beta n\, e^{i\phi} + (1-\beta)n\, e^{i\phi'}| < n$ almost surely, and for $\beta = 1$, $S_{c^\star} = n\, t_{g(a^\star)}$ exactly. Nor can a bound of order $\beta^{-2}\log K$ hold for arbitrary summaries: one bit per stratum and task, $\mathbf{1}\{T_c(x) \ge \beta\}$, suffices on the needle ensemble. The dimension lower bound is therefore a statement about additive sketches read through linear probes, which is the regime in which SIEVE and its upper bound operate.

A.11 Proof of Theorem 2

Upper bound. Let $d \ge 512A\beta^{-2}\log\frac{4K(L+1)}{\delta}$ and $8v_c\Delta_c \le \beta$ for all c. By Prop. 2, SIEVE is successful on every feasible population, and by Lem. 3 its expected cost is at most $\lceil 16\beta^{-1}\log(2/\delta)\rceil$, independent of K and N. For fixed A, L, δ and $K \ge 2$, $\log\frac{4K(L+1)}{\delta} = O(\log K)$, so $d = O(\beta^{-2}\log K)$ suffices.

Lower bound. Fix $\alpha \in (0,1)$ and work in the readout-noise model at the calibrated scale of Rem. 3. If $d \le \alpha^2\beta^{-2}\log K$, then equation 7 holds, and by Lem. 9 every successful procedure has average, and hence worst-case, expected cost at least $\frac{(1-\delta)^2}{12\beta}K^{(1-\alpha)^2}$ on the needle ensemble once $K^{(1-\alpha)^2} \ge 6\beta/(1-\delta)$. This grows polynomially in K.

Conclusion. Suppose that, in the readout-noise model, a dimension $d(K)$ admits a procedure with expected cost at most a constant C for all K. Then the lower bound forces $d(K) > \alpha^2\beta^{-2}\log K$ as soon as $K^{(1-\alpha)^2} > \max\{6\beta/(1-\delta),\, 12\beta C/(1-\delta)^2\}$, while the upper bound shows that $d = 512A\beta^{-2}\log\frac{4K(L+1)}{\delta}$ suffices. Hence $d^* = \Theta(\beta^{-2}\log K)$ for coordinators that read each summary through one noisy linear readout. The upper bound holds for every population; the needle ensemble is the population that forces the lower bound. □