TEST-TIME SCALING VIA BUDGETED MULTI-ATTRIBUTE VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Verifying LLM-generated answers under a shared computational budget requires jointly deciding which candidates to inspect and which verification attributes to evaluate. We formulate this problem as multi-attribute good-arm identification under a global budget: each candidate is an arm evaluated along several costly attributes, and the goal is to certify as many candidates as possible whose mean scores exceed the prescribed thresholds on all attributes. We propose BMA-GAI, an algorithm that combines cost-aware arm selection with adaptive sampling of attributes. Every observation serves both to guide adaptive allocation and to support anytime-valid certification, which removes the need for a separate confirmation stage. We establish an asymptotic coverage guarantee for BMA-GAI and derive a matching information-theoretic converse that characterizes the intrinsic complexity of the problem, thereby proving that BMA-GAI is first-order optimal away from critical budget levels. Experiments on synthetic benchmarks and an LLM answer-verification task show that BMA-GAI allocates the verification budget more efficiently and certifies more high-quality candidates than competing methods.

1 INTRODUCTION

Test-time scaling has emerged as an effective approach to improving the performance of large language models (LLMs) by allocating additional computation at inference time. Existing methods scale test-time computation through repeated sampling and aggregation (Wang et al., 2023; Brown et al., 2024), search over alternative reasoning trajectories (Yao et al., 2023), iterative refinement with feedback (Madaan et al., 2023; Shinn et al., 2023), and verifier-guided selection among candidate solutions (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2024). Recent studies further show that reasoning performance can be improved by explicitly controlling test-time computation (Snell et al., 2025; Muennighoff et al., 2025), and that the marginal benefit of additional compute depends strongly on both input difficulty and inference strategy (Snell et al., 2025). These observations have motivated adaptive allocation across queries, either by predicting the marginal value of additional computation (Damani et al., 2025) or by casting cross-query allocation as a bandit problem (Zuo & Zhu, 2026; Wang et al., 2025a).

Verification plays a central role in this allocation problem. In candidate-based reasoning, additional generations are useful only if promising responses can be reliably distinguished from incorrect ones (Cobbe et al., 2021; Lightman et al., 2024). Recent work has therefore explored multiple specialized verifiers (Lifshitz et al., 2025) and adaptive allocation of verification computation (Zhong et al., 2026; Singhi et al., 2025). These developments suggest viewing verification as a test-time resource whose allocation directly determines how many candidate answers can be reliably certified.

In many applications, however, the validity of an answer cannot be adequately summarized by a single verification score. Different verifiers may assess complementary requirements, such as numerical consistency, logical validity, and other task-specific criteria (Lifshitz et al., 2025; Kwok et al., 2026). Moreover, verifier calls can differ substantially in both cost and statistical difficulty, which makes the allocation of verification effort nontrivial in its own right (Kanarios et al., 2024). This heterogeneity enables cost-aware screening: inexpensive verifiers can provide early evidence that a candidate is unlikely to satisfy all criteria, so that expensive verification can be reserved for candidates that remain plausible. At the same time, certification is conjunctive: a candidate can be accepted only after sufficient evidence has been accumulated for *every* required attribute, as in multi-attribute constrained

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.

identification (Sravani et al., 2024). Under a shared verification budget, the learner must therefore decide both *which candidate* to verify and *which attribute* to inspect next. This leads to our central question: *how should a limited verification budget be allocated across candidates and verification attributes to reliably certify as many candidates as possible?*

We formalize this setting as *multi-attribute good-arm identification under a global verification budget*. Each fixed query–answer pair is modeled as an arm with multiple separately sampleable verification attributes, where evaluating an attribute incurs a known cost and yields a noisy score. An arm is certified only if all of its attributes meet their prescribed thresholds, and the learner aims to certify as many arms as possible within a shared hard budget while controlling the probability of any erroneous certification. This formulation builds on three lines of work: good-arm identification, which seeks arms whose means exceed a prescribed threshold (Kano et al., 2019); multi-attribute constrained bandits with separately observable feasibility criteria (Sravani et al., 2024); and cost-aware pure exploration with heterogeneous sampling costs (Kanarios et al., 2024). In contrast to these settings, we study a budgeted *coverage* objective: the learner allocates verification effort under a shared budget so as to maximize the number of reliably certified arms. The achievable coverage is therefore not fixed in advance, but is jointly determined by the budget and the arms' certification costs.

To answer this question, we propose BMA-GAI, an adaptive allocation algorithm that simultaneously learns which candidates are inexpensive to certify and how to distribute verification effort across their unresolved attributes. Our main contributions are summarized as follows:

- We formulate *multi-attribute good-arm identification under a global verification budget*, in which each candidate must satisfy multiple separately observable criteria with heterogeneous evaluation costs, and the goal is to maximize reliable certification coverage.
- We develop BMA-GAI, which adaptively allocates verification effort across candidates and their unresolved attributes via cost-aware selection, and we establish an achievable coverage guarantee showing that it reliably certifies as many good candidates as the available budget permits.
- We characterize the fundamental complexity of reliable coverage: certifying ℓ good arms requires an aggregate complexity of $T^*_\ell(\mu)$ (up to a $\log(1/\delta)$ factor), which is determined by the ℓ least costly good arms. A matching information-theoretic converse shows that BMA-GAI attains this frontier to first order away from critical budget thresholds.
- We evaluate BMA-GAI on synthetic benchmarks and an LLM answer-verification task, demonstrating the benefits of adaptively allocating verification effort across candidates and heterogeneous verification attributes under limited budgets.

2 Related Work

Test-Time Scaling for LLMs. Test-time scaling has become a prominent paradigm for improving LLM reasoning without modifying model parameters. One class of methods increases inference-time computation by generating multiple reasoning trajectories and then aggregating or selecting among them. Representative approaches include self-consistency and its adaptive or voting-based variants (Wang et al., 2023; Chen et al., 2023; Aggarwal et al., 2023; Xue et al., 2023), as well as repeated-sampling and Best-of-N strategies that select promising responses from larger candidate pools (Brown et al., 2024; Snell et al., 2025; Wang et al., 2025b). A second line of work spends additional computation on searching over or iteratively refining reasoning trajectories, including Tree-of-Thoughts-style search (Yao et al., 2023), self-refinement (Madaan et al., 2023), and feedback-driven iterative reasoning (Shinn et al., 2023). A third class devotes inference-time computation to evaluating generated candidates, using verification signals to assess complete solutions or intermediate reasoning steps (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2024). More broadly, recent studies show that the benefit of additional test-time computation depends strongly on the task, input difficulty, and inference strategy (Snell et al., 2025; Muennighoff et al., 2025), motivating a shift from uniformly increasing inference compute toward allocating it selectively.

Recent work has increasingly explored *adaptive* test-time scaling, including input-adaptive allocation of decoding computation (Damani et al., 2025), bandit-based budget allocation across queries (Zuo & Zhu, 2026; Wang et al., 2025a), and adaptive verification through multiple verifiers or dynamic allocation of verification effort (Lifshitz et al., 2025; Zhong et al., 2026; Singhi et al., 2025). In

contrast, we study how to allocate a verification budget across a fixed set of candidate answers and multiple verification attributes, with the objective of maximizing reliable certification coverage.

Good-Arm Identification and Cost-Aware Pure Exploration. A broad literature on bandit pure exploration studies threshold-based identification (Locatelli et al., 2016; Kano et al., 2019; Cho et al., 2025; Jiang et al., 2025), problems with multiple correct answers (Degenne & Koolen, 2019), structured feasibility constraints (Sravani et al., 2024; Lardy et al., 2025; Cai & Kandasamy, 2026; Mukherjee & Moharir, 2026), and heterogeneous sampling costs or resource constraints (Kanarios et al., 2024; Li & Cheung, 2024). Thresholding bandits classify arms according to whether their mean rewards exceed a prescribed threshold (Locatelli et al., 2016), whereas good-arm identification (GAI) aims to sequentially output satisfactory arms as soon as they can be certified (Kano et al., 2019). Multiple-answer pure exploration allows several outputs to be simultaneously correct and studies how sampling should be allocated among competing valid answers (Degenne & Koolen, 2019). More recent work extends these ideas to anytime-valid and nonparametric GAI (Cho et al., 2025) and to settings with multiple threshold constraints (Jiang et al., 2025).

A complementary line of work incorporates explicit resource constraints and heterogeneous sampling costs into pure exploration (Kanarios et al., 2024; Li & Cheung, 2024; Sravani et al., 2024; Lardy et al., 2025; Cai & Kandasamy, 2026; Mukherjee & Moharir, 2026). Cost-aware best-arm identification accounts for arm-dependent sampling costs (Kanarios et al., 2024), while resource-constrained formulations impose global limits on the resources consumed during identification (Li & Cheung, 2024). Closest to our feedback model, constrained grouped bandits consider arms composed of separately observable attributes subject to conjunctive feasibility requirements (Sravani et al., 2024), and subsequent work studies related constrained and fixed-budget variants (Lardy et al., 2025; Cai & Kandasamy, 2026; Mukherjee & Moharir, 2026). Our setting combines these two lines of research but pursues a different goal: certifying as many arms as possible under a shared hard budget, where each certified arm must satisfy all prescribed attribute-wise thresholds.

3 PRELIMINARIES

We formulate test-time answer verification as a *multi-attribute bandit problem* under a global verification budget $B > 0$. There are Q candidate arms indexed by $[Q] \coloneqq \{1, \dots, Q\}$. In the LLM setting, each arm q corresponds to a fixed query–answer pair, and a_q denotes the associated candidate answer. Each arm has M separately sampleable attributes indexed by $[M] \coloneqq \{1, \dots, M\}$, and sampling attribute m of arm q amounts to pulling the arm–attribute pair (q, m). Each attribute corresponds to a fixed verification procedure (a verifier) that assesses a particular aspect of answer validity. A fresh invocation of verifier m on candidate answer a_q incurs a known cost $c_m > 0$ and returns a stochastic verification score with unknown mean $\mu_{q,m} \in [0, 1]$, normalized so that larger values indicate stronger evidence of validity. Each attribute m is associated with a prescribed acceptance threshold $\xi_m \in (0, 1)$.

Budgeted bandit interaction. At each round t, the learner adaptively selects an arm–attribute pair $(q_t, m_t) \in [Q] \times [M]$ and observes a stochastic verification score Y_t. Let $\mathcal{F}_{t-1}$ denote the history available before round t, including the learner's internal randomization, so that (q_t, m_t) is $\mathcal{F}_{t-1}$-measurable. We assume that each new observation satisfies

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = \mu_{q_t,m_t}, \quad \mathbb{E}[\exp(\lambda(Y_t - \mu_{q_t,m_t})) \mid \mathcal{F}_{t-1}] \le \exp\left(\frac{\lambda^2}{2}\right), \quad \forall \lambda \in \mathbb{R}. \tag{1}$$

Thus, repeated pulls of the same arm–attribute pair yield fresh verifier observations, and the learner may adapt its sampling decisions to all previously observed evidence. Since each pull of attribute m incurs cost c_m, the cumulative cost after t observations is

$$C_t = \sum_{s=1}^{t} c_{m_s}.$$

The learner is given a global verification budget $B > 0$. Because verification costs may vary across attributes, the budget constrains the total incurred cost rather than merely the number of verifier calls. The learner must therefore decide adaptively both which arm–attribute pairs to sample and how to distribute the available budget among them.

Good arms. An arm is deemed good if all of its attribute means meet their respective acceptance thresholds. Accordingly, the set of good arms is

$$\mathcal{G} = \{q \in [Q] : \mu_{q,m} \geq \xi_m,\ \forall m \in [M]\}.$$

In LLM verification, $q \in \mathcal{G}$ means that the candidate answer a_q satisfies all verification criteria. This definition induces a fundamental asymmetry: violating a single threshold suffices to disqualify an arm, whereas certification requires evidence that *all* thresholds are met. Consequently, the verification budget should be allocated adaptively, rather than uniformly, across candidate answers and attributes. To characterize both the algorithmic guarantee and the intrinsic problem complexity, we assume $\mu_{q,m} \neq \xi_m$ for all $(q, m) \in [Q] \times [M]$.

Budgeted reliable coverage. For a prescribed confidence level $\delta \in (0, 1)$, a policy π consists of an adaptive sampling rule, a stopping time T_{stop} taking nonnegative integer values, and an $\mathcal{F}_{T_{\text{stop}}}$-measurable output set $\widehat{\mathcal{S}}_\pi(B) \subseteq [Q]$. The policy may depend on the budget B, the confidence level δ, and the known costs and thresholds. Our objective is to maximize the expected number of returned arms subject to family-wise reliability and a hard verification budget:

$$\sup_\pi \mathbb{E}_{\boldsymbol{\mu},\pi}\left[\left|\widehat{\mathcal{S}}_\pi(B)\right|\right] \qquad \text{s.t.}\, \mathbb{P}_{\boldsymbol{\mu},\pi}\left(\widehat{\mathcal{S}}_\pi(B) \subseteq \mathcal{G}\right) \geq 1 - \delta, \quad C_{T_{\text{stop}}} \leq B.$$

Here, $\boldsymbol{\mu} = (\mu_{q,m} : (q, m) \in [Q] \times [M])$ denotes the matrix of unknown means. The objective is thus to maximize reliable coverage under a fixed cumulative-cost budget. Since the achievable coverage depends on both the budget and the difficulty of the underlying instance, effective verification calls for adaptive allocation across arms and attributes.

4 Algorithm

We now present BMA-GAI, an adaptive allocation algorithm for multi-attribute good-arm identification under a global verification budget. The algorithm maintains a set of certified arms and adaptively allocates verifier calls among the remaining arms. Each observation serves both to guide subsequent allocation and to accumulate evidence for certification, and the algorithm terminates once all arms are certified or no further verifier call is affordable. The full procedure is given in Algorithm 1.

Let $D = QM$. For each $t = 0, 1, 2, \ldots$, let $\mathcal{S}_t$, $\mathcal{Q}_t$, $\mathcal{M}_t$, and C_t denote, respectively, the set of certified arms, the set of active arms, the set of currently affordable attributes, and the cumulative verification cost after t observations. For each arm–attribute pair (q, m), let $N_{q,m}(t)$, $\widehat{\mu}_{q,m}(t)$, and $L_{q,m}(t)$ denote its sample count, empirical mean, and lower certificate. Time indices are suppressed in Algorithm 1 for notational simplicity. BMA-GAI initializes $\mathcal{S}_0 = \varnothing$, $\mathcal{Q}_0 = [Q]$, $C_0 = 0$, and $\mathcal{M}_0 = \{m \in [M] : C_0 + c_m \leq B\}$, and sets $N_{q,m}(0) = \widehat{\mu}_{q,m}(0) = L_{q,m}(0) = 0$ for each (q, m).

To describe the sampling rule, we first specify the certification state maintained by the algorithm. For each pair (q, m), the lower certificate after t observations is

$$L_{q,m}(t) = \max\left\{L_{q,m}(t-1), \widehat{\mu}_{q,m}(t) - \sqrt{\frac{2}{N_{q,m}(t)} \log \frac{4DN_{q,m}^2(t)}{\delta}}\right\}, \quad L_{q,m}(0) = 0, \tag{2}$$

whenever $N_{q,m}(t) > 0$; otherwise, it remains unchanged. Based on these certificates, the set of unresolved attributes of arm q is

$$\mathcal{U}_q(t) = \{m \in [M] : L_{q,m}(t) < \xi_m\}.$$

Arm q is certified once $\mathcal{U}_q(t) = \varnothing$, i.e., once the lower certificate of every attribute reaches its threshold. Certified arms are then removed from the active set, while all remaining arms continue to participate in sampling.

Sampling decision. Given the current certification state, BMA-GAI selects the next arm–attribute pair (q_{t+1}, m_{t+1}) from the active arms and currently affordable attributes, which includes three cases:

(i) Sparse exploration. The algorithm first checks whether some currently available arm–attribute pair is under-explored. Specifically, if

$$\min_{q \in \mathcal{Q}_t,\, m \in \mathcal{M}_t} N_{q,m}(t) < \frac{\sqrt{t+1}}{D}, \tag{3}$$

Algorithm 1 BMA-GAI: Budgeted Multi-Attribute Good-Arm Identification

Input: Arms $\mathcal{A} = \{a_q\}_{q=1}^{Q}$; thresholds $\{\xi_m\}_{m=1}^{M}$; costs $\{c_m\}_{m=1}^{M}$; budget $B > 0$; $\delta \in (0, 1/2)$

Initialize $\mathcal{S} \leftarrow \varnothing$, $\mathcal{Q} \leftarrow [Q]$, $t \leftarrow 0$, $C \leftarrow 0$, $\mathcal{M} \leftarrow \{m \in [M] : C + c_m \le B\}$
Initialize $(N_{q,m}, \widehat{\mu}_{q,m}, L_{q,m}) \leftarrow (0, 0, 0)$, $\forall (q, m) \in [Q] \times [M]$
while $\mathcal{Q} \neq \varnothing$ **and** $\mathcal{M} \neq \varnothing$ **do**
  if Eq. (3) holds **then**
    Choose $(q, m) \in \arg\min_{q \in \mathcal{Q},\, m \in \mathcal{M}} N_{q,m}$
  else
    Compute ℓ, u, H^-, H^+ using Eqs. (4)–(5)
    if $\min_{q \in \mathcal{Q}} H_q^+ = \infty$ **then**
      Choose $(q, m) \in \arg\min_{q \in \mathcal{Q},\, m \in \mathcal{M}} N_{q,m}$
    else
      Choose q using Eq. (6) and m using Eq. (7)
    end if
  end if
  Obtain one fresh observation Y from (q, m)
  $C \leftarrow C + c_m$, $t \leftarrow t + 1$
  Update $N_{q,m}$, $\widehat{\mu}_{q,m}$, and $L_{q,m}$ using Eq. (2)
  $\mathcal{C} \leftarrow \{q \in \mathcal{Q} : L_{q,m} \ge \xi_m,\ \forall m \in [M]\}$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{C}$, $\mathcal{Q} \leftarrow \mathcal{Q} \setminus \mathcal{C}$
  $\mathcal{M} \leftarrow \{m \in [M] : C + c_m \le B\}$
end while
return $\widehat{\mathcal{S}} = \mathcal{S}$; abstain on all remaining arms

it selects a least-sampled pair in $\mathcal{Q}_t \times \mathcal{M}_t$ as (q_{t+1}, m_{t+1}). This sparse exploration prevents the allocation rule from over-relying on early estimates and ensures that all currently available arm–attribute pairs continue to be explored.

(ii) Cost-aware arm screening. If the sparse-exploration condition is not triggered, the algorithm uses the available observations to assess which active arms are potentially inexpensive to certify. It first constructs the allocation intervals

$$\ell_{q,m}(t) = \widehat{\mu}_{q,m}(t) - \sqrt{\frac{8\log(t+1)}{N_{q,m}(t)}}, \quad u_{q,m}(t) = \widehat{\mu}_{q,m}(t) + \sqrt{\frac{8\log(t+1)}{N_{q,m}(t)}}. \tag{4}$$

When $N_{q,m}(t) = 0$, we set $\ell_{q,m}(t) = 0$ and $u_{q,m}(t) = 1$, based on the known range $\mu_{q,m} \in [0, 1]$. These intervals are used only for allocation; they do not enter the certification rule and introduce no additional confidence parameter.

Let $\ell_q(t) = (\ell_{q,m}(t))_{m=1}^{M}$ and $u_q(t) = (u_{q,m}(t))_{m=1}^{M}$ denote the vectors of interval endpoints. For a vector $z \in \mathbb{R}^M$, define the certification-cost surrogate

$$H(z) = \begin{cases} \displaystyle\sum_{m=1}^{M} \frac{2c_m}{(z_m - \xi_m)^2}, & z_m > \xi_m,\ \forall m \in [M], \\ +\infty, & \text{otherwise.} \end{cases}$$

The algorithm then computes the operational cost indices

$$\big(H_q^-(t), H_q^+(t)\big) = \begin{cases} \big(H(u_q(t)), H(\ell_q(t))\big), & \mathcal{U}_q(t) \cap \mathcal{M}_t \neq \varnothing, \\ (+\infty, +\infty), & \mathcal{U}_q(t) \cap \mathcal{M}_t = \varnothing. \end{cases} \tag{5}$$

Here, $H_q^-(t)$ and $H_q^+(t)$ are optimistic and conservative estimates, respectively, of the verification cost required to certify arm q. The value $+\infty$ in the second case serves only as an allocation-feasibility device that prevents target selection from choosing an arm with no currently affordable unresolved attribute; it neither constitutes a permanent rejection nor implies that the arm has infinite intrinsic statistical complexity. In the budget-free reference process used in our analysis (Appendix B), $\mathcal{M}_t = [M]$ for all t, so the operational and raw indices coincide for every active uncertified arm.

If $\min_{q \in \mathcal{Q}_t} H_q^+(t) = \infty$, then either the current observations do not yet yield a finite conservative cost estimate for any active arm, or no active arm has an affordable unresolved attribute. In this case, BMA-GAI falls back to sparse exploration and selects a least-sampled pair in $\mathcal{Q}_t \times \mathcal{M}_t$.

(iii) Stable target and attribute allocation. Once at least one active arm admits a finite conservative cost estimate, the algorithm constructs the set of potentially least-cost arms and selects the one with the smallest index as the target:

$$q_{t+1} = \min \mathcal{P}_t, \quad \mathcal{P}_t = \left\{ q \in \mathcal{Q}_t : H_q^-(t) \le \min_{i \in \mathcal{Q}_t} H_i^+(t) \right\}. \tag{6}$$

Here, $\mathcal{P}_t$ consists of the active arms that cannot yet be ruled out as the least costly to certify. Selecting the smallest index provides a stable tie-breaking rule and avoids unnecessary target switching caused by small fluctuations in the empirical estimates.

Given the target arm q_{t+1}, the algorithm then allocates verification effort across its unresolved attributes. For each $m \in \mathcal{U}_{q_{t+1}}(t)$, define the optimistic sample requirement $\widetilde{v}_{q_{t+1},m}(t) = 2/(u_{q_{t+1},m}(t) - \xi_m)^2$. By construction of the operational index, $\mathcal{U}_{q_{t+1}}(t) \cap \mathcal{M}_t$ is nonempty. Moreover, finiteness of $H^-_{q_{t+1}}(t)$ implies $u_{q_{t+1},m}(t) > \xi_m$ for every unresolved attribute m, so the following selection rule is well defined:

$$m_{t+1} \in \arg \min_{m \in \mathcal{U}_{q_{t+1}}(t) \cap \mathcal{M}_t} \frac{N_{q_{t+1},m}(t)}{\widetilde{v}_{q_{t+1},m}(t)}. \tag{7}$$

Thus, verification is directed to the unresolved attribute whose current sample count is smallest relative to its requirement. Verifier costs enter target-arm selection through $H_q^-(t)$ and $H_q^+(t)$, whereas Eq. (7) balances statistical evidence across the unresolved attributes of the target arm.

Observation and certification. After selecting (q_{t+1}, m_{t+1}), BMA-GAI obtains a fresh verifier observation Y_{t+1} and updates the cumulative cost to $C_{t+1} = C_t + c_{m_{t+1}}$. Based on the updated certificates, the algorithm identifies the set of newly certified arms

$$\mathcal{C}_{t+1} = \{ q \in \mathcal{Q}_t : L_{q,m}(t+1) \ge \xi_m, \ \forall m \in [M] \}.$$

The newly certified arms are added to the certified set and removed from the active set:

$$\mathcal{S}_{t+1} = \mathcal{S}_t \cup \mathcal{C}_{t+1}, \quad \mathcal{Q}_{t+1} = \mathcal{Q}_t \setminus \mathcal{C}_{t+1}.$$

Thus, $\mathcal{S}_t$ records all arms certified within the first t observations and grows monotonically. In contrast, an arm that does not yet meet the certification condition remains in $\mathcal{Q}_t$ and may receive further verification in subsequent rounds. Certification is therefore the only irreversible decision made by the algorithm; unreturned arms are never permanently rejected.

Budget update and termination. After each observation, BMA-GAI also updates the set of affordable attributes according to the remaining budget:

$$\mathcal{M}_{t+1} = \{ m \in [M] : C_{t+1} + c_m \le B \}.$$

In other words, an attribute is excluded from further consideration once its cost exceeds the remaining budget. The next round then proceeds with the active set $\mathcal{Q}_{t+1}$ and the affordable set $\mathcal{M}_{t+1}$. This sampling–certification cycle continues as long as $\mathcal{Q}_t \neq \varnothing$ and $\mathcal{M}_t \neq \varnothing$. The algorithm stops at time T_{stop} when either all arms have been certified ($\mathcal{Q}_{T_{\text{stop}}} = \varnothing$) or no further verifier call is affordable ($\mathcal{M}_{T_{\text{stop}}} = \varnothing$). It then returns the accumulated certified set $\widehat{\mathcal{S}}(B) = \mathcal{S}_{T_{\text{stop}}}$ and abstains on the remaining active arms. Hence, the number of returned answers is not fixed in advance, but is determined adaptively by the available budget and the statistical difficulty of the instance.

5 Theoretical Guarantees

In this section, we characterize both the performance of BMA-GAI and the intrinsic difficulty of multi-attribute certification under a shared verification budget. Together, matching achievability and converse bounds identify the first-order optimal coverage.

Proposition 1 (Anytime reliability and hard-budget feasibility)**.** *For any fixed budget $B > 0$,* BMA-GAI *satisfies $C_{T_{stop}} \le B$ almost surely and*

$$\mathbb{P}_\mu(\mathcal{S}_t \subseteq \mathcal{G} \text{ for all } t \le T_{stop}) \ge 1 - \delta. \tag{8}$$

Proposition 1 thus provides a time-uniform certification guarantee: with probability at least $1-\delta$, every arm certified at any time before termination is good, despite the adaptive allocation across arms and attributes. Moreover, the cumulative verification cost never exceeds the prescribed budget, so reliability is attained without violating the hard budget constraint.

To characterize both the algorithmic guarantee and the intrinsic problem complexity, for $q \in \mathcal{G}$, define

$$\Delta_{q,m} = \mu_{q,m} - \xi_m, \quad h_{q,m} = \frac{2c_m}{\Delta_{q,m}^2}, \quad h_q = \sum_{m=1}^{M} h_{q,m}.$$

Here, $h_{q,m}$ quantifies the complexity of certifying attribute m of arm q, accounting for both the evaluation cost c_m and the threshold gap $\Delta_{q,m}$, and h_q aggregates this complexity over all attributes. Building on these arm-wise complexities, define, for $1 \le \ell \le |\mathcal{G}|$,

$$T_\ell^*(\mu) = \min_{\substack{\mathcal{S}\subseteq\mathcal{G} \\ |\mathcal{S}|=\ell}} \sum_{q\in\mathcal{S}} h_q, \quad T_0^*(\mu) = 0. \tag{9}$$

Thus, $T_\ell^*(\mu)$ is the minimum aggregate complexity required to certify any ℓ good arms, attained by the ℓ good arms with the smallest h_q. The following theorem shows that this benchmark directly governs the coverage achieved by BMA-GAI under a global budget.

Theorem 1 (Achievable coverage under a global budget)**.** *Fix an analysis slack $\varepsilon \in (0,1)$ and a normalized budget $b > 0$, and define $\underline{K}_\varepsilon(b) = \max\{\ell \in \{0,\dots,|\mathcal{G}|\} : (1+\varepsilon)T_\ell^*(\mu) < b\}$. Then, with $B_\delta = b\log(1/\delta)$,* BMA-GAI *satisfies*

$$\mathbb{P}_\mu\Big(\widehat{\mathcal{S}}(B_\delta) \subseteq \mathcal{G}, \quad |\widehat{\mathcal{S}}(B_\delta)| \ge \underline{K}_\varepsilon(b)\Big) \longrightarrow 1 \quad \text{as } \delta \downarrow 0.$$

Theorem 1 provides an asymptotic performance guarantee for BMA-GAI at normalized budget b: with probability tending to one, the algorithm returns only good arms and certifies at least $\underline{K}_\varepsilon(b)$ of them. In other words, BMA-GAI asymptotically attains every coverage level ℓ whose aggregate certification complexity satisfies $(1+\varepsilon)T_\ell^*(\mu) < b$. The proof first bounds the certification costs in a budget-free reference process and then couples the finite-budget algorithm with this process for as long as every attribute remains affordable.

We next show that, under Gaussian observations, the same quantity $T_\ell^*(\mu)$ also characterizes the intrinsic complexity of the problem. Specifically, consider independent unit-variance observations $\mathcal{N}(\mu_{q,m}, 1)$, and let $\Pi_\delta(B)$ denote the class of policies that are uniformly δ-reliable over this Gaussian model class and satisfy the hard budget B on every admissible instance.

Theorem 2 (Information-theoretic coverage limit)**.** *Fix $\varepsilon \in (0,1)$ and $b > 0$, and define $\overline{K}_\varepsilon(b) = \max\{\ell \in \{0,\dots,|\mathcal{G}|\} : (1-\varepsilon)T_\ell^*(\mu) \le b\}$. Then, with $B_\delta = b\log(1/\delta)$,*

$$\sup_{\pi\in\Pi_\delta(B_\delta)} \mathbb{P}_{\mu,\pi}\left(|\widehat{\mathcal{S}}_\pi(B_\delta)| > \overline{K}_\varepsilon(b)\right) \longrightarrow 0 \quad \text{as } \delta \downarrow 0.$$

Theorem 2 shows that no uniformly reliable policy can asymptotically certify more than $\overline{K}_\varepsilon(b)$ arms with nonvanishing probability. Together with Theorem 1, this identifies $T_\ell^*(\mu)$ as the first-order complexity governing achievable coverage: BMA-GAI attains every coverage level whose complexity lies below the budget, whereas no uniformly reliable policy can exceed the corresponding information-theoretic limit.

Combining the achievability and converse results yields the following first-order optimality guarantee.

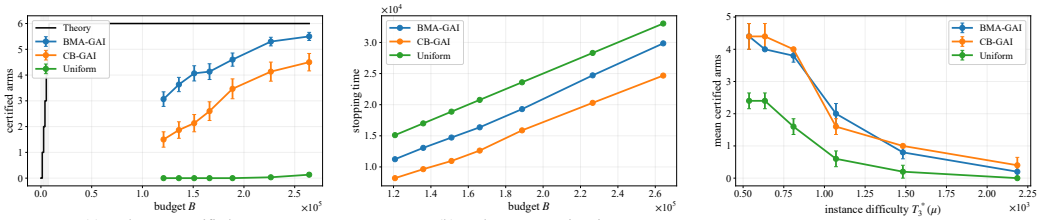
Corollary 1 (First-order optimal coverage)**.** *Under the Gaussian model, fix $b > 0$ with $b \notin \{T_\ell^*(\mu) : 1 \le \ell \le |\mathcal{G}|\}$, and define $K(b) = \max\{\ell \in \{0,\dots,|\mathcal{G}|\} : T_\ell^*(\mu) \le b\}$. Then, for $B_\delta = b\log(1/\delta)$,*

$$\mathbb{P}_\mu\left(\widehat{\mathcal{S}}(B_\delta) \subseteq \mathcal{G}, \quad |\widehat{\mathcal{S}}(B_\delta)| = K(b)\right) \longrightarrow 1 \quad \text{as } \delta \downarrow 0.$$

Moreover,

$$\lim_{\delta\downarrow 0} \mathbb{E}_\mu[|\widehat{\mathcal{S}}(B_\delta)|] = \lim_{\delta\downarrow 0} \sup_{\pi\in\Pi_\delta(B_\delta)} \mathbb{E}_{\mu,\pi}[|\widehat{\mathcal{S}}_\pi(B_\delta)|] = K(b).$$

Thus, $K(b)$ is the maximum reliable coverage permitted by the problem complexity at normalized budget b, and BMA-GAI asymptotically attains it; that is, the number of good arms certified by BMA-GAI matches the information-theoretic limit of the instance to first order. Critical budgets $b = T_\ell^*(\mu)$ are excluded to avoid boundary cases in which lower-order terms determine whether the ℓth arm can be certified.



(a) Budget vs. certified arms (b) Budget vs. stopping time (c) $T_3^*(\mu)$ vs. certified arms

Figure 1: Synthetic experiments. (a) Reliable coverage compared with the characteristic-complexity staircase; (b) the corresponding stopping time; (c) coverage as the characteristic difficulty $T_3^*(\mu)$ varies across problem instances. Error bars indicate standard errors.

6 Experiments

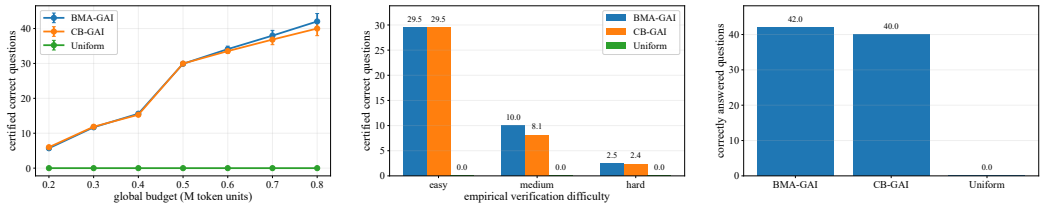
We evaluate BMA-GAI on controlled synthetic instances and on a verification task derived from GSM8K (Cobbe et al., 2021). The synthetic experiments examine reliable coverage under a fixed budget and its dependence on the intrinsic certification complexity, and compare the results with the complexity predicted by our theory. The GSM8K experiment evaluates whether BMA-GAI can effectively allocate a limited verification budget across cheap, medium-cost, and expensive verifiers.

Baselines. We compare BMA-GAI with two baselines. Uniform spreads the budget evenly across active arm–attribute pairs and applies the same certification rule as BMA-GAI, thereby measuring the benefit of adaptive allocation. CB-GAI is a cost-blind variant of good-arm identification that uses the same confidence intervals and threshold-certification rule but ignores verifier costs when selecting arm–attribute pairs, thereby isolating the benefit of cost-aware allocation.

6.1 Synthetic experiments

Experimental setup. All synthetic experiments use Gaussian observations $Y_{q,m} \sim \mathcal{N}(\mu_{q,m}, 1)$ with thresholds $\xi_m = 0.5$, $Q = 12$ arms, and $M = 4$ attributes. Figures 1(a,b) use a heterogeneous-cost instance with $|\mathcal{G}| = 6$ and costs $c = (1, 2, 4, 25)$. The instance is constructed so that some bad arms appear plausible under the cheap attributes and can be ruled out only with evidence from the expensive attribute, whereas good arms can be certified through more cost-efficient attributes. We fix $\delta = 0.02$ and vary the budget $B = b \log(1/\delta)$ over seven levels, with normalized budgets $b \in \{16, 18, 20, 22, 25, 30, 35\} \times T^*_{|\mathcal{G}|}(\mu)$. All methods use the same anytime certification radius and differ only in their allocation rules. Figure 1(c) uses a second family of instances with costs $c = (1, 2, 4, 8)$, in which the good-arm gaps are scaled to produce different values of $T_3^*(\mu)$; for this experiment, we fix $B = 90000$ and $\delta = 0.02$. We report the mean number of certified good arms and the mean stopping time, together with standard errors, over 30 independent repetitions.

Results. Figure 1 shows a consistent advantage of BMA-GAI over both baselines. As shown in Figure 1(a), BMA-GAI achieves substantially higher reliable coverage across a wide range of budgets, whereas the coverage of Uniform grows slowly because it wastes observations on arm–attribute pairs that are not critical for certification. Although CB-GAI benefits from adaptive sampling, it ignores attribute-dependent costs and may therefore spend excessive budget on arms that look promising after cheap observations but ultimately require expensive attributes to be certified or ruled out. By jointly accounting for estimated gaps and verifier costs, BMA-GAI prioritizes arms whose remaining certificates are both statistically plausible and cheap to complete, and thus uses the budget more effectively. The staircase pattern in Figure 1(a) also agrees with the characteristic-complexity thresholds: an additional arm becomes reliably certifiable only once the budget exceeds its corresponding complexity level. Figure 1(b) further confirms that the gain stems from more efficient sampling rather than a weaker stopping criterion, since all methods use the same anytime certificate. Finally, Figure 1(c) shows that the coverage of all methods decreases as $T_3^*(\mu)$ increases, supporting its role as an intrinsic measure of instance difficulty; nevertheless, the persistent advantage of BMA-GAI indicates that cost-aware allocation remains beneficial across difficulty levels. Additional experimental results on budget allocation and the interaction between verification cost and statistical difficulty are provided in Appendix A.



(a) Budget vs. answered questions (b) Coverage by difficulty (c) Correctly answered questions

Figure 2: GSM8K experiments. (a) Reliable coverage as the global budget varies; (b) certified correct answers stratified by empirical verification difficulty; (c) final number of correctly answered questions at the largest budget. Error bars indicate standard errors.

6.2 Real-world experiments

Experimental setup. We construct a real-world instance from 100 GSM8K test questions. For each question, we generate one fixed candidate solution with Qwen3-8B and treat each question–candidate pair as an arm. Each arm has three verifier attributes: the cheap verifier checks whether the solution contains a parseable final answer; the medium verifier solves the question once with Qwen3-8B and compares its answer with the candidate answer; and the expert verifier repeats this procedure five times and compares the majority answer with the candidate answer. Each verifier call returns a binary accept/reject outcome together with its token cost, consistent with our multi-attribute bandit model; during each run, verifier observations are replayed by sampling with replacement from precomputed outcomes. We set the acceptance threshold to $\xi_m = 0.5$ for all three attributes. The GSM8K ground-truth answers are never accessible to any algorithm during verification; they are used only after each run to assess whether the certified candidates are correct.

Results. Figure 2 demonstrates the advantage of adaptive and cost-aware verification on GSM8K. As shown in Figure 2(a), Uniform certifies almost no answers under limited budgets: by spreading verification effort evenly across questions and verifiers, it leaves most questions with insufficient evidence to meet the certification criterion. In contrast, BMA-GAI and CB-GAI adaptively concentrate verification on questions that are more likely to be certified. Figure 2(b) further shows that the adaptive methods certify easy and medium questions first before spending heavily on hard ones, which improves budget efficiency by avoiding excessive effort on questions that remain difficult to resolve. The gap between BMA-GAI and CB-GAI stems mainly from cost awareness: BMA-GAI explicitly accounts for heterogeneous verifier costs and therefore favors verifier calls that yield certification evidence more cost-efficiently. Finally, because certification is performed without access to the ground truth and relies on imperfect verifiers, a certified answer is not necessarily correct. Figure 2(c) therefore compares the certified candidates with the gold answers after each run to measure the actual useful coverage, on which BMA-GAI again achieves the best overall performance.

7 Conclusion and Future Work

This paper studied multi-attribute good-arm identification under a global verification budget, motivated by the allocation of test-time verification computation across LLM-generated candidate answers. We proposed BMA-GAI, which jointly selects candidate answers and verification attributes via cost-aware adaptive allocation while providing anytime-valid certification under a hard budget. Our analysis identified the aggregate certification complexity $T^*_\ell(\mu)$ as the fundamental quantity governing reliable coverage. We established an achievable coverage guarantee together with a matching information-theoretic converse, showing that BMA-GAI attains the optimal coverage to first order away from critical budget thresholds. Experiments on synthetic instances and GSM8K further demonstrated that adaptive, cost-aware verification uses a shared budget more efficiently than uniform or cost-blind allocation.

Several directions remain open. An important one is to jointly allocate computation between candidate generation and verification. It would also be interesting to study richer verification models with unknown or time-varying evaluation costs, and to derive sharper finite-budget guarantees.

AI USE STATEMENT

We used generative AI tools to assist with drafting and revising the manuscript, checking technical consistency, and developing experimental code. In the GSM8K experiments, Qwen3-8B was used to generate the candidate solutions and the solution samples underlying the verification procedures. The authors take full responsibility for the problem formulation, theoretical claims, experimental methodology, and reported results, including all text, code, and figures prepared with AI assistance.

REPRODUCIBILITY STATEMENT

The problem formulation and the BMA-GAI algorithm are presented in Sections 3 and 4, and proofs of the main theoretical results are given in Appendix B. Section 6 describes the synthetic Gaussian experiments and the GSM8K-derived verification task, together with the baselines and evaluation metrics. The GSM8K experiments use one fixed candidate answer per question and replay observations by sampling with replacement from precomputed verifier outcomes with recorded token costs. To facilitate reproduction, we will publicly release the code and materials upon acceptance, including synthetic instance generators, implementations of the algorithm and baselines, GSM8K question identifiers, candidate solutions, verifier outcomes and costs, experimental configurations and random seeds, raw trial outputs, result summaries, and plotting scripts.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12375–12396, 2023.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL https://arxiv.org/abs/2407.21787.

Ting Cai and Kirthevasan Kandasamy. Constrained best arm identification with tests for feasibility. *Proceedings of the 40th AAAI Conference on Artificial Intelligence*, pp. 19808–19816, 2026.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation, 2023. URL https://arxiv.org/abs/2311.17311.

Brian M. Cho, Dominik Meier, Kyra Gan, and Nathan Kallus. Reward maximization for pure exploration: Minimax optimal good arm identification for nonparametric multi-armed bandits. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pp. 2251–2259, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. Learning how hard to think: Input-adaptive allocation of LM computation. In *The 13th International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6qUUgw9bAZ.

Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems 32*, pp. 14564–14573, 2019.

Xuanke Jiang, Sherief Hashima, Kohei Hatano, and Eiji Takimoto. Multi-thresholding good arm identification with bandit feedback. In *Proceedings of the 17th Asian Conference on Machine Learning*, pp. 49–64, 2025.

Kellen Kanarios, Qining Zhang, and Lei Ying. Cost aware best arm identification. *Reinforcement Learning Journal*, 4:1533–1545, 2024.

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. doi: 10.1007/s10994-019-05784-4.

Jacky Kwok, Shulu Li, Pranav Atreya, Yuejiang Liu, Yixing Jiang, Chelsea Finn, Marco Pavone, Ion Stoica, and Azalia Mirhoseini. LLM-as-a-verifier: A general-purpose verification framework, 2026. URL https://arxiv.org/abs/2607.05391.

Tyron Lardy, Christina Katsimerou, and Wouter M. Koolen. Constrained best arm identification. In *Advances in Neural Information Processing Systems 38*, pp. 111559–111585, 2025.

Zitian Li and Wang Chi Cheung. Best arm identification with resource constraints. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 253–261, 2024.

Shalev Lifshitz, Sheila A. McIlraith, and Yilun Du. Multi-agent verification: Scaling test-time compute with multiple verifiers, 2025. URL https://arxiv.org/abs/2502.20379.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The 12th International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1690–1698, 2016.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36*, pp. 46534–46594, 2023.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, 2025.

Raunak Mukherjee and Sharayu Moharir. Fixed-budget constrained best arm identification in grouped bandits, 2026. URL https://arxiv.org/abs/2603.04007.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36*, pp. 8634–8652, 2023.

Nishad Singhi, Hritik Bansal, Arian Hosseini, Aditya Grover, Kai-Wei Chang, Marcus Rohrbach, and Anna Rohrbach. When to solve, when to verify: Compute-optimal problem solving and generative verification for LLM reasoning. In *Conference on Language Modeling*, 2025.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The 13th International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Malyala Preethi Sravani, Sakshi Heda, and Sharayu Moharir. Constrained best arm identification in grouped bandits. In *2024 International Conference on Signal Processing and Communications (SPCOM)*, pp. 25–29, 2024.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback, 2022. URL https://arxiv.org/abs/2211.14275.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arık. DynScaling: Efficient verifier-free inference scaling via dynamic and integrated sampling, 2025a. URL https://arxiv.org/abs/2506.16043.

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 *The 11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yiming Wang, Pei Zhang, Siyuan Huang, Baosong Yang, Zhuosheng Zhang, Fei Huang, and Rui Wang. Sampling-efficient test-time scaling: Self-estimating the best-of-n sampling in early decoding. In *Advances in Neural Information Processing Systems 38*, pp. 180020–180057, 2025b.

Mingfeng Xue, Dayiheng Liu, Wenqiang Lei, Xingzhang Ren, Baosong Yang, Jun Xie, Yidan Zhang, Dezhong Peng, and Jiancheng Lv. Dynamic voting for efficient reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3085–3104, 2023.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36*, pp. 11809–11822, 2023.

Jianyuan Zhong, Zeju Li, Zhijian Xu, Xiangyu Wen, Kezhi Li, and Qiang Xu. Solve-detect-verify: Inference-time scaling with flexible generative verifier. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47377–47401, 2026.

Bowen Zuo and Yinglun Zhu. Strategic scaling of test-time compute: A bandit learning approach. In *The 14th International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0mNnINd2z5.

A Additional Experiments

We present two additional synthetic diagnostics that examine how the verification budget is distributed across arms and how cost heterogeneity interacts with statistical difficulty. Both use the same three methods as Section 6. The arm categories shown in the figures are used only for analysis and are never revealed to any algorithm.

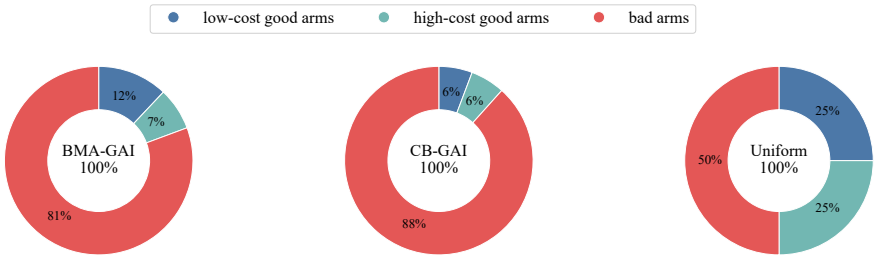
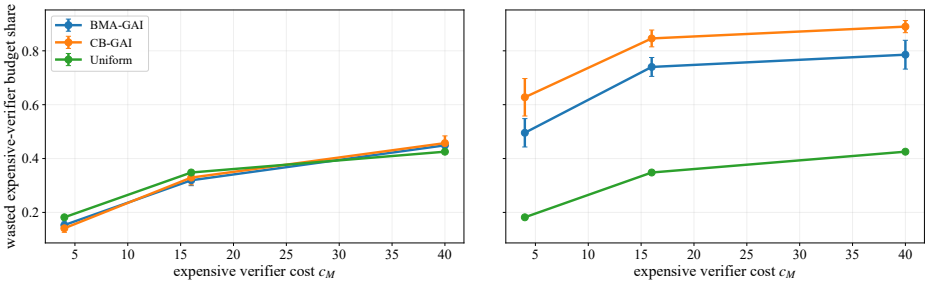


Figure 3: Final budget allocation on the expensive-decoy synthetic instance. Each circular panel corresponds to one algorithm after the budget is exhausted, and its three sectors show the budget shares spent on low-certification-cost good arms, high-certification-cost good arms, and bad arms.

Final budget allocation by arm type. Figure 3 shows the final budget allocation across three arm types on a stress-test instance with $Q = 12$ arms, $M = 4$ attributes, costs $c = (1, 2, 4, 25)$, and $|\mathcal{G}| = 6$. The bad arms appear statistically promising on the inexpensive attributes and fail only on the costly one. The good arms are split into low- and high-certification-cost groups according to their characteristic costs h_q. The total budget is set to $12T^*_{|\mathcal{G}|}(\mu)\log(1/\delta)$ with $\delta = 0.02$, and each panel averages results over eight independent repetitions.

The resulting allocation patterns reveal clear differences among the methods. Uniform spreads its budget nearly evenly across good and bad arms and therefore cannot accumulate enough evidence to certify good arms in this budget regime. Both adaptive methods instead concentrate their budgets on statistically promising arms. However, CB-GAI spends a larger fraction of its budget on the bad decoy arms because it ignores verifier costs when selecting targets. In contrast, BMA-GAI devotes more budget to low-certification-cost good arms, reflecting its preference for arms whose certificates are both statistically plausible and cheap to complete.



Bad arms fail cheap verifier | Bad arms fail expensive verifier

Figure 4: Cost–difficulty interaction. The vertical axis shows the cost of expensive-verifier calls on bad arms as a fraction of the total budget; error bars denote standard errors over eight repetitions. The left and right panels differ in which attribute reveals that a bad arm is bad.

Interaction between cost and difficulty. This experiment examines whether the benefit of cost-aware allocation depends on *where* the evidence for rejecting a bad arm is concentrated. Both instance families contain $Q = 12$ arms and $M = 4$ attributes, with six good arms sharing the same mean vectors and verifier costs $(1, 2, 4, c_M)$, where $c_M \in \{4, 16, 40\}$. In both families, every bad arm falls below the threshold by the same small gap. In the left panel, this gap occurs on the cheapest attribute, so bad arms can be rejected with inexpensive evidence. In the right panel, the gap occurs only on the

fourth attribute: the cheaper attributes make the bad arms look promising, and rejecting them requires costly observations. For each instance, we set $B = 12T^*_{|\mathcal{G}|}(\mu)\log(1/\delta)$ with $\delta = 0.02$ and average the results over eight independent repetitions. The reported quantity is the fraction of the total budget B spent on applying the fourth verifier to truly bad arms.

When bad arms can be rejected with the cheapest verifier, BMA-GAI and CB-GAI spend similar fractions of the budget on expensive verification of bad arms, even as the cost of the fourth verifier increases. In contrast, when bad arms can be ruled out only through the expensive verifier, CB-GAI consistently spends a larger fraction of the budget on costly bad-arm observations than BMA-GAI, and the gap widens as c_M increases. This more efficient allocation also translates into substantially more certified good arms for BMA-GAI in the high-cost regime. Although Uniform may spend a smaller fraction on expensive verification of bad arms, it fails to certify good arms under the same budget, showing that a small fraction alone does not indicate effective allocation. Together, the two panels show that cost awareness is most beneficial when costly verification coincides with the main source of identification difficulty.

B Proofs for BMA-GAI

Throughout the appendix, write

$$D = QM, \quad c_{\min} = \min_m c_m, \quad c_{\max} = \max_m c_m.$$

In both coverage theorems, $b > 0$ is fixed and $B_\delta = b\log(1/\delta)$. Define the certificate radius

$$r(n,\delta) = \sqrt{\frac{2}{n}\log\frac{4Dn^2}{\delta}}, \quad n \geq 1.$$

For any execution, let $\sigma_{q,m}(n) = \inf\{t : N_{q,m}(t) = n\}$, with $\inf\varnothing = \infty$. On $\{\sigma_{q,m}(n) < \infty\}$, write $\bar{X}_{q,m}(n)$ for the empirical mean of its first n observations of pair (q,m). Unlike $\widehat{\mu}_{q,m}(t)$, this notation is indexed by the local sample count. The sub-Gaussian proofs never require an empirical mean at an unattained local count.

B.1 Proof of Proposition 1

Proof. Fix $B > 0$. For a pair (q,m), let $I_s^{q,m} = \mathbf{1}\{(q_s, m_s) = (q,m)\}$ and

$$W_{q,m}(t) = \sum_{s=1}^{t} I_s^{q,m}(Y_s - \mu_{q,m}).$$

For every fixed $\lambda \in \mathbb{R}$, the conditional moment assumption implies that

$$Z_{q,m}^{\lambda}(t) = \exp\left(\lambda W_{q,m}(t) - \frac{\lambda^2}{2}N_{q,m}(t)\right)$$

is a nonnegative supermartingale starting at one; after termination, we keep this process constant. Optional stopping at $\sigma_{q,m}(n) \wedge T$, followed by Fatou's lemma as $T \to \infty$, gives

$$\mathbb{E}_\mu\left[\mathbf{1}\{\sigma_{q,m}(n) < \infty\}\exp\left(\lambda n(\bar{X}_{q,m}(n) - \mu_{q,m}) - \frac{\lambda^2 n}{2}\right)\right] \leq 1.$$

Optimizing the resulting exponential bound at $\lambda = x$ or $-x$ yields, for either choice of sign,

$$\mathbb{P}_\mu\big(\sigma_{q,m}(n) < \infty,\ \pm(\bar{X}_{q,m}(n) - \mu_{q,m}) \geq x\big) \leq e^{-nx^2/2}, \quad x > 0. \tag{10}$$

This argument applies to any admissible adaptive execution, not only to the finite-budget algorithm. Taking $x = r(n,\delta)$ and combining the two tails gives

$$\begin{aligned}&\mathbb{P}_\mu\big(\sigma_{q,m}(n) < \infty,\ |\bar{X}_{q,m}(n) - \mu_{q,m}| > r(n,\delta)\big)\\ &\quad\leq 2e^{-nr^2(n,\delta)/2} = \frac{\delta}{2Dn^2}.\end{aligned}$$

Note that the event includes attainment of the local count; the probability is not conditioned on attainment. A union bound over all pairs and counts therefore gives

$$\mathbb{P}_\mu\big(\exists(q,m), n \geq 1 :\ \sigma_{q,m}(n) < \infty,\ |\bar{X}_{q,m}(n) - \mu_{q,m}| > r(n,\delta)\big) \leq D \sum_{n=1}^{\infty} \frac{\delta}{2Dn^2} = \frac{\pi^2}{12}\delta \leq \delta.$$

On the complementary event, every attained raw lower bound is at most $\mu_{q,m}$. Since $\mu_{q,m} \geq 0$, the running maximum defining $L_{q,m}(t)$, including its initial value zero, also satisfies $L_{q,m}(t) \leq \mu_{q,m}$. Thus, certification of q implies $\mu_{q,m} \geq \xi_m$ for every m, and hence $q \in \mathcal{G}$. This holds simultaneously at all times up to termination and proves Eq. (8).

Every sampling branch selects an attribute from the current set $\mathcal{M}_t = \{m : C_t + c_m \leq B\}$. In the ordinary branch, the operational-index convention in Eq. (5) ensures that the selected target has a nonempty affordable unresolved set. Consequently, every executed observation satisfies $C_{t+1} = C_t + c_{m_{t+1}} \leq B$. Induction gives $C_{T_{\text{stop}}} \leq B$ almost surely, and positivity of the costs gives $T_{\text{stop}} \leq \lfloor B/c_{\min} \rfloor$.

Storing counts, empirical means, and certificates uses $O(QM)$ memory. Direct scans for exploration, cost indices, selection, and certification use $O(QM)$ operations per observation, excluding verifier execution. □

B.2 Proof of Theorem 1

Throughout this subsection, we assume only the conditional 1-sub-Gaussian observation model and a fixed nonboundary mean matrix, i.e., $\mu_{q,m} \neq \xi_m$ for all (q,m). Neither Gaussianity nor independence across observations is required.

Reference process. Consider the action rule obtained from Algorithm 1 by setting $\mathcal{M}_t = [M]$ at every decision, with all other rules unchanged. We refer to an execution of this rule as a budget-free reference process. For every active arm, $\mathcal{U}_q(t) \neq \varnothing$, so its operational cost indices agree with $H(u_q(t))$ and $H(\ell_q(t))$. Define

$$\tau_\ell(\delta) = \inf\{t : |\mathcal{S}_t| \geq \ell\}, \quad \mathcal{C}_\ell(\delta) = C_{\tau_\ell(\delta)},$$

with $\mathcal{C}_\ell(\delta) = \infty$ when $\tau_\ell(\delta) = \infty$. All probability estimates below hold uniformly over the conditional observation laws satisfying Eq. (1) with these means. In particular, they do not require different confidence levels to share a common observation stream.

Lemma 1 (Sparse-exploration count)**.** *Let $F(t)$ count the forced-exploration observations among the first t observations of the reference process. Then*

$$F(t) \leq D \left\lceil \frac{\sqrt{t}}{D} \right\rceil \leq \sqrt{t} + D. \tag{11}$$

For $t \geq t_D := 16(D+1)^2$, every arm still active after observation t satisfies

$$N_{q,m}(t) \geq \left\lfloor \frac{\sqrt{t/2}}{D} \right\rfloor \geq \frac{\sqrt{t}}{3D}, \quad m \in [M].$$

Proof. When a pair is selected by forced exploration at observation $s \leq t$, its pre-draw count is smaller than $\sqrt{s}/D \leq \sqrt{t}/D$. Its count increases after every such selection, so it can be selected this way at most $\lceil \sqrt{t}/D \rceil$ times. Summing over all D pairs proves Eq. (11).

Suppose an arm remains active at time t but one of its pairs has count smaller than $\lfloor \sqrt{t/2}/D \rfloor$. The arm was active at all earlier times because certification is irreversible. At every observation s in the last half of the first t observations, that pair's pre-draw count is below $\sqrt{s}/D$. Each such observation would therefore be forced exploration, implying $F(t) \geq \lfloor t/2 \rfloor$. For $t \geq t_D$, this contradicts $\lfloor t/2 \rfloor > \sqrt{t} + D$. The final floor bound follows from $t \geq 16(D+1)^2$. □

Lemma 2 (High-probability stabilization)**.** *There exists a deterministic $t_*(\mu) < \infty$ such that, for every integer $u \geq 2$, there is an event $\mathcal{A}_u$ with*

$$\mathbb{P}_\mu(\mathcal{A}_u^c) \leq \frac{D}{(u-1)^2},$$

on which, at every non-exploration decision with $t \geq \max\{u, t_*(\mu)\}$ at which some good arm is still active, $\min_{q \in \mathcal{Q}_t} H_q^+(t) < \infty$ and

$$\mathcal{P}_t = \arg\min_{q \in \mathcal{Q}_t \cap \mathcal{G}} h_q.$$

Proof. Let $\mathcal{A}_u$ be the event that

$$|\bar{X}_{q,m}(n) - \mu_{q,m}| \leq \sqrt{\frac{8\log(t+1)}{n}}$$

for every integer $t \geq u$, every pair, and every $1 \leq n \leq t$ satisfying $\sigma_{q,m}(n) \leq t$. By Eq. (10),

$$\mathbb{P}_\mu(\mathcal{A}_u^c) \leq 2D \sum_{t=u}^{\infty} \frac{t}{(t+1)^4} \leq 2D \sum_{t=u}^{\infty} t^{-3} \leq \frac{D}{(u-1)^2}.$$

On $\mathcal{A}_u$, every active pair with $t \geq \max\{u, t_D\}$ has allocation radius at most

$$e(t) = \sqrt{\frac{24D\log(t+1)}{\sqrt{t}}} \longrightarrow 0,$$

by Lemma 1. Its allocation interval contains the mean, and both endpoints lie within $2e(t)$ of the mean.

Since all threshold gaps are nonzero, eventually every active good arm has all lower endpoints strictly above the thresholds, whereas every active bad arm has an upper endpoint strictly below its threshold on at least one attribute. Thus bad arms have $H_q^- = H_q^+ = \infty$, and good arms satisfy $H_q^- \leq h_q \leq H_q^+$, with both endpoints converging uniformly to h_q.

If there are distinct good-arm complexities, let $\rho > 0$ be their smallest positive separation. Choose a deterministic time large enough such that both cost endpoints of every good arm differ from h_q by less than $\rho/3$. An arm with complexity larger than the smallest remaining complexity then has its optimistic cost strictly above a conservative cost of a smallest-complexity arm, and is excluded from $\mathcal{P}_t$. Every arm tied at the smallest remaining complexity $h_{\min}$ remains in $\mathcal{P}_t$, since $H_q^- \leq h_{\min} \leq \min_{i \in \mathcal{Q}_t} H_i^+$. If all good-arm complexities are equal, only this latter argument is needed. These bounds apply to every active subset. Taking $t_*(\mu)$ to include these deterministic requirements proves the claim. Fixed-index tie-breaking selects the first remaining good arm in the order (h_q, q). □

For a good pair, define the first crossing among its attained counts:

$$n_{q,m}^\dagger(\delta) = \inf\left\{n \geq 1 : \sigma_{q,m}(n) < \infty,\ \bar{X}_{q,m}(n) - r(n,\delta) \geq \xi_m\right\}.$$

The value is ∞ if the set is empty.

Lemma 3 (High-probability one-attribute certification)**.** *For each good pair and fixed $\eta > 0$,*

$$\mathbb{P}_\mu\left(n_{q,m}^\dagger(\delta) > \frac{2(1+\eta)}{\Delta_{q,m}^2}\log(1/\delta)\right) \longrightarrow 0.$$

Consequently,

$$\mathbb{P}_\mu\left(n_{q,m}^\dagger(\delta) \leq \frac{2(1+\eta)}{\Delta_{q,m}^2}\log(1/\delta), \quad \forall q \in \mathcal{G},\ m \in [M]\right) \longrightarrow 1. \tag{12}$$

Proof. Fix a good pair and abbreviate its mean, threshold, and gap by ν, ξ, and $\Delta = \nu - \xi > 0$. Set

$$n_\delta = \left\lfloor \frac{2(1+\eta)}{\Delta^2}\log(1/\delta) \right\rfloor.$$

For sufficiently small δ, $n_\delta \geq 1$ and

$$r(n_\delta, \delta) \longrightarrow \frac{\Delta}{\sqrt{1+\eta}} < \Delta.$$

Choose a fixed $a_\eta > 0$ such that, for all sufficiently small δ, $r(n_\delta, \delta) \leq \Delta - a_\eta$.

On the event $\{n^\dagger(\delta) > n_\delta\}$, the local count n_δ must be attained. Otherwise, this attribute never crosses its threshold, its arm is never certified, and the reference process keeps that arm active forever; the count lower bound in Lemma 1 then contradicts nonattainment of n_δ. At the attained count, absence of a crossing gives

$$\bar{X}(n_\delta) - \nu < r(n_\delta, \delta) - \Delta \leq -a_\eta.$$

Therefore Eq. (10) implies

$$\mathbb{P}_\mu(n^\dagger(\delta) > n_\delta) \leq e^{-n_\delta a_\eta^2/2} \longrightarrow 0.$$

Since $n^\dagger$ is integer-valued or infinite, the event on the left is exactly $\{n^\dagger > 2(1+\eta)\log(1/\delta)/\Delta^2\}$. A union bound over the finitely many good pairs proves the simultaneous statement. □

Proposition 2 (Reference-process completion upper bound)**.** *For every fixed $\alpha \in (0,1)$, under the conditional sub-Gaussian model,*

$$\mathbb{P}_\mu(\mathcal{C}_\ell(\delta) \leq (1+\alpha) T_\ell^*(\mu)\log(1/\delta), \quad \forall \ell \in \{1, \ldots, |\mathcal{G}|\}) \longrightarrow 1. \tag{13}$$

The convergence is uniform over admissible conditional observation laws with the fixed mean matrix.

Proof. If $|\mathcal{G}| = 0$, the statement is vacuous. Otherwise set $\eta = \alpha/4$ and $u_\delta = \max\{2, \lceil [\log(1/\delta)]^{1/3} \rceil\}$. Order the good arms by (h_q, q) as $q_{(1)}, \ldots, q_{(|\mathcal{G}|)}$, and set $S_\ell^* = \{q_{(1)}, \ldots, q_{(\ell)}\}$. Then $\sum_{q \in S_\ell^*} h_q = T_\ell^*(\mu)$ by Eq. (9). Let $\mathcal{B}_\delta$ be the event in Eq. (12). The preceding lemmas give

$$\mathbb{P}_\mu(\mathcal{A}_{u_\delta} \cap \mathcal{B}_\delta) \longrightarrow 1.$$

For sufficiently small δ, $u_\delta \geq t_*(\mu)$. On this intersection, define

$$A_{\ell,\delta}^\dagger = \sum_{q \in S_\ell^*} \sum_{m=1}^{M} c_m n_{q,m}^\dagger(\delta).$$

Then

$$A_{\ell,\delta}^\dagger \leq (1+\eta) T_\ell^*(\mu)\log(1/\delta), \quad 1 \leq \ell \leq |\mathcal{G}|.$$

We charge the first u_δ observations at cost at most $c_{\max} u_\delta$. Subsequently, while fewer than ℓ outputs have occurred, every ordinary target belongs to S_ℓ^*. Indeed, stable selection of a good arm with rank larger than ℓ would require all of the first ℓ good arms to have already been certified. This argument remains valid even if erroneous certifications occur earlier, because arms are removed only through certification.

An ordinary draw samples an unresolved attribute, whose pre-draw count must be smaller than $n_{q,m}^\dagger(\delta)$. Once this count has been reached, the running certificate stays above threshold. The total post-u_δ ordinary cost is therefore at most $A_{\ell,\delta}^\dagger$. Stabilization also excludes the fallback branch while a good arm remains active. All other draws are forced exploration. Thus every prefix up to and including the observation producing the ℓth output satisfies

$$C(t) \leq A_{\ell,\delta}^\dagger + c_{\max}(u_\delta + D) + c_{\max}\sqrt{t}.$$

Writing

$$W_{\ell,\delta} = A_{\ell,\delta}^\dagger + c_{\max}(u_\delta + D), \quad \kappa = \frac{c_{\max}}{\sqrt{c_{\min}}},$$

and using $C(t) \geq c_{\min} t$, we obtain $C(t) \leq W_{\ell,\delta} + \kappa\sqrt{C(t)}$. Solving this inequality gives

$$\begin{aligned} C(t) &\leq W_{\ell,\delta} + \frac{\kappa^2}{2} + \frac{\kappa}{2}\sqrt{\kappa^2 + 4W_{\ell,\delta}} \\ &\leq W_{\ell,\delta} + \kappa^2 + \kappa\sqrt{W_{\ell,\delta}}. \end{aligned} \tag{14}$$

The ℓth output must occur on the event under consideration; otherwise, a good arm remains active forever, so the reference process keeps sampling and $C(t) \geq c_{\min} t \to \infty$, contradicting Eq. (14).

Because $u_\delta = o(\log(1/\delta))$, the deterministic bounds on this event imply, simultaneously for the finitely many ℓ,

$$W_{\ell,\delta} \leq (1+\eta) T_\ell^*(\mu) \log(1/\delta) + o(\log(1/\delta)), \quad \kappa^2 + \kappa\sqrt{W_{\ell,\delta}} = o(\log(1/\delta)).$$

Since $\eta < \alpha$ and every $T_\ell^*(\mu) > 0$ for $\ell \geq 1$, Eq. (13) follows. The bounds on $\mathbb{P}_\mu(\mathcal{A}_{u_\delta}^c)$ and $\mathbb{P}_\mu(\mathcal{B}_\delta^c)$ depend only on δ and the fixed instance parameters, not on the particular conditional observation laws, which proves the asserted uniformity. □

Lemma 4 (Budget coupling)**.** *For each fixed (B, δ), a finite-budget execution and an admissible reference execution can be coupled to agree on every reference prefix whose cumulative cost is at most $B - c_{\max}$. In particular,*

$$\mathcal{C}_\ell(\delta) \leq B - c_{\max} \quad \Longrightarrow \quad |\widehat{\mathcal{S}}(B)| \geq \ell, \quad 1 \leq \ell \leq |\mathcal{G}|.$$

Proof. If $B < c_{\max}$, the displayed implication is vacuous. Otherwise, start both executions with the same action–observation history and fixed tie-breaking rules. Whenever their histories agree and $C_t \leq B - c_{\max}$, every attribute is affordable because

$$C_t + c_m \leq C_t + c_{\max} \leq B, \quad m \in [M].$$

Thus $\mathcal{M}_t = [M]$ in the finite-budget execution. Its exploration, cost-index, target-selection, and attribute-selection rules agree with those of the reference process. Feeding both executions the same next observation, drawn from the finite-budget experiment's conditional feedback law, keeps their counts, means, certificates, and certified sets in agreement.

We continue synchronously while the cost condition holds. Once it fails, the reference process can be extended using any conditional observation law satisfying Eq. (1) with the same means; for example, it may receive the selected pair's mean deterministically. This is only an auxiliary probabilistic construction, not a change to the algorithm. It requires neither independent streams nor a common sample path for different budgets. Proposition 2 applies uniformly to such continuations.

Induction proves agreement on all the stated prefixes. If $\mathcal{C}_\ell(\delta) \leq B - c_{\max}$, the finite-budget execution has made the same ℓ certifications by that point. Subsequent budget-aware decisions cannot revoke them, proving the implication. □

Proof of Theorem 1. Fix $b > 0$ and $\varepsilon \in (0,1)$, and write $k = \underline{K}_\varepsilon(b)$. If $k = 0$, reliability alone proves the claim. Suppose $k \geq 1$. By definition,

$$d_b := b - (1+\varepsilon) T_k^*(\mu) > 0.$$

For sufficiently small δ, $d_b \log(1/\delta) \geq c_{\max}$, so

$$(1+\varepsilon) T_k^*(\mu) \log(1/\delta) \leq b \log(1/\delta) - c_{\max} = B_\delta - c_{\max}. \tag{15}$$

Construct the coupled reference process from Lemma 4. By Proposition 2, with $\alpha = \varepsilon$,

$$\mathbb{P}_\mu(\mathcal{C}_k(\delta) \leq (1+\varepsilon) T_k^*(\mu) \log(1/\delta)) \longrightarrow 1.$$

Its uniformity allows the auxiliary continuation to depend on (B_δ, δ). Combining this estimate with Eq. (15) and the coupling lemma gives $\mathbb{P}_\mu(|\widehat{\mathcal{S}}(B_\delta)| \geq k) \to 1$. Finally,

$$\begin{aligned} &\mathbb{P}_\mu\Big(\widehat{\mathcal{S}}(B_\delta) \subseteq \mathcal{G},\ |\widehat{\mathcal{S}}(B_\delta)| \geq k\Big) \\ &\quad \geq 1 - \delta - \mathbb{P}_\mu(\mathcal{C}_k(\delta) > (1+\varepsilon) T_k^*(\mu) \log(1/\delta)) \longrightarrow 1, \end{aligned}$$

by Proposition 1. This proves the theorem. □

B.3 Proof of Theorem 2

In this subsection, observations are independent unit-variance Gaussian streams. They may be realized as independent arrays $X_{q,m,n} \sim \mathcal{N}(\mu_{q,m}, 1)$, revealing $X_{q,m,n}$ at the nth draw of (q,m). This Gaussian construction is not used in the sub-Gaussian achievability proof.

Variational characterization. For nonempty $S \subseteq \mathcal{G}$, let $\mathrm{Alt}(S) = \{\lambda \in [0,1]^{Q\times M} : S \not\subseteq \mathcal{G}(\lambda)\}$ and define

$$\Gamma(S) = \max_{w\in\Delta_D} \inf_{\lambda\in\mathrm{Alt}(S)} \sum_{q,m} \frac{w_{q,m}}{c_m} \frac{(\mu_{q,m}-\lambda_{q,m})^2}{2},$$

where Δ_D is the probability simplex of cost fractions.

Lemma 5 (Variational characterization)**.** *For every nonempty $S \subseteq \mathcal{G}$,*

$$\Gamma(S) = \frac{1}{\sum_{q\in S} h_q}. \tag{16}$$

Consequently, for $1 \le \ell \le |\mathcal{G}|$,

$$\left[\max_{S\subseteq\mathcal{G}:\,|S|=\ell} \Gamma(S)\right]^{-1} = T^*_\ell(\mu).$$

Proof. Every alternative invalidating S has some $q \in S, m \in [M]$ with $\lambda_{q,m} < \xi_m$. The information contributed by this coordinate is at least $w_{q,m}\Delta^2_{q,m}/(2c_m)$. Conversely, changing just one such coordinate and approaching ξ_m from below attains this value as an infimum. Since $\xi_m \in (0,1)$, these alternatives remain in the mean domain. Therefore, for each fixed w,

$$\inf_{\lambda\in\mathrm{Alt}(S)} \sum_{q,m} \frac{w_{q,m}}{c_m} \frac{(\mu_{q,m}-\lambda_{q,m})^2}{2} = \min_{q\in S, m\in[M]} \frac{w_{q,m}}{h_{q,m}}.$$

If the minimum equals z, then $w_{q,m} \ge z h_{q,m}$ on S. Summing gives $1 \ge z\sum_{q\in S} h_q$. Equality is achieved by $w_{q,m} = h_{q,m}/\sum_{i\in S} h_i$ on S, with zero weight elsewhere. This proves Eq. (16). Maximizing over size-ℓ sets selects the ℓ smallest good-arm complexities. The corresponding within-arm cost fractions are $h_{q,m}/h_q$. □

Lemma 6 (Necessary evidence for returning a good arm)**.** *Fix $\varepsilon \in (0,1)$. For all sufficiently small δ, uniformly over $B > 0$, $\pi \in \Pi_\delta(B)$, $q \in \mathcal{G}$, and $m \in [M]$,*

$$\mathbb{P}_{\mu,\pi}\left(q \in \widehat{\mathcal{S}}_\pi(B),\ N_{q,m}(T_{stop}) < \frac{2(1-\varepsilon)}{\Delta^2_{q,m}} \log(1/\delta)\right) \le 2\delta^{\varepsilon^2/64}.$$

The required upper bound on δ depends only on the fixed instance and ε.

Proof. Fix a good pair, put $\nu = \mu_{q,m}$, $\xi = \xi_m$, and $\Delta = \nu - \xi > 0$. Choose

$$u = \min\left\{\frac{\xi}{2},\ \Delta\left(\sqrt{1+\varepsilon/4}-1\right)\right\} > 0.$$

Then $u < \xi$ and

$$\frac{(\Delta+u)^2}{\Delta^2} \le 1 + \frac{\varepsilon}{4}. \tag{17}$$

Define an alternative mean matrix λ by changing only $\lambda_{q,m} = \xi - u$. Arm q is bad on this admissible alternative, so uniform reliability gives

$$\mathbb{P}_{\lambda,\pi}(q \in \widehat{\mathcal{S}}_\pi(B)) \le \delta.$$

Write $d = \Delta + u$ and $n_\delta = \lfloor 2(1-\varepsilon)\log(1/\delta)/\Delta^2 \rfloor$. For small enough δ, $n_\delta \ge 1$. Under the true instance, let $S_k = \sum_{i=1}^k (X_{q,m,i} - \nu)$, with $S_0 = 0$. The local log-likelihood ratio is

$$Z_k = dS_k + \frac{d^2 k}{2}.$$

The hard budget bounds the stopping time by $\lfloor B/c_{\min} \rfloor$ under both instances. We pad the transcript after stopping to this common finite horizon. The policy's action, randomization, and output kernels

are identical under both instances, and observations of all other pairs have identical laws. Thus the full transcript likelihood ratio is $\exp(Z_N)$, where $N = N_{q,m}(T_{\text{stop}})$.

Let $A = \{q \in \widehat{\mathcal{S}}_\pi(B),\ N \le n_\delta\}$ and $x_\delta = (1-\varepsilon/2)\log(1/\delta)$. Splitting at $Z_N = x_\delta$ gives

$$\begin{aligned}\mathbb{P}_{\mu,\pi}(A) &\le e^{x_\delta}\mathbb{P}_{\lambda,\pi}(A) + \mathbb{P}_{\mu,\pi}(A, Z_N > x_\delta)\\ &\le \delta^{\varepsilon/2} + \mathbb{P}_\mu\left(\max_{0\le k\le n_\delta} Z_k > x_\delta\right).\end{aligned} \tag{18}$$

By Eq. (17),

$$\frac{d^2 n_\delta}{2} \le (1-\varepsilon)(1+\varepsilon/4)\log(1/\delta) \le (1-3\varepsilon/4)\log(1/\delta).$$

Consequently,

$$\mathbb{P}_\mu\left(\max_{k\le n_\delta} Z_k > x_\delta\right) \le \mathbb{P}_\mu\left(\max_{k\le n_\delta} S_k > \frac{\varepsilon\log(1/\delta)}{4d}\right).$$

For a Gaussian random walk, stopping $\exp(\theta S_k - \theta^2 k/2)$ at its first crossing of $S_k > a$ before n, and optimizing at $\theta = a/n$, gives

$$\mathbb{P}\left(\max_{0\le k\le n} S_k > a\right) \le e^{-a^2/(2n)}.$$

Since $d^2 n_\delta \le 2\log(1/\delta)$, it follows that

$$\begin{aligned}\mathbb{P}_\mu\left(\max_{k\le n_\delta} Z_k > x_\delta\right) &\le \exp\left(-\frac{\varepsilon^2\log^2(1/\delta)}{32 d^2 n_\delta}\right)\\ &\le \delta^{\varepsilon^2/64}.\end{aligned}$$

Combining with Eq. (18) and using $\delta^{\varepsilon/2} \le \delta^{\varepsilon^2/64}$ proves the bound for A. Finally, $\{N < 2(1-\varepsilon)\log(1/\delta)/\Delta^2\} \subseteq \{N \le n_\delta\}$. All constants are independent of the policy and budget; finiteness of the set of good pairs gives a common sufficiently small δ. □

Proof of Theorem 2. Fix $b > 0$ and $\varepsilon \in (0,1)$, let $B_\delta = b\log(1/\delta)$, and write $\bar{k} = \overline{K}_\varepsilon(b)$. For any $\pi \in \Pi_\delta(B_\delta)$, let

$$\mathcal{B}_\pi = \{\widehat{\mathcal{S}}_\pi(B_\delta) \not\subseteq \mathcal{G}\}$$

and

$$\mathcal{R}_\pi = \bigcup_{q\in\mathcal{G}, m\in[M]} \left\{q \in \widehat{\mathcal{S}}_\pi(B_\delta),\ N_{q,m}(T_{\text{stop}}) < \frac{2(1-\varepsilon)}{\Delta_{q,m}^2}\log(1/\delta)\right\}.$$

Reliability and Lemma 6 imply that

$$\mathbb{P}_{\mu,\pi}(\mathcal{B}_\pi) \le \delta, \quad \mathbb{P}_{\mu,\pi}(\mathcal{R}_\pi) \le 2|\mathcal{G}|M\delta^{\varepsilon^2/64}.$$

On the complement of these two events, all returned arms are good and every required attribute has at least its stated number of observations. If k arms are returned, then $k \le |\mathcal{G}|$ and

$$\begin{aligned}b\log(1/\delta) = B_\delta &\ge \sum_{q\in\widehat{\mathcal{S}}_\pi(B_\delta)} \sum_{m=1}^{M} c_m N_{q,m}(T_{\text{stop}})\\ &\ge (1-\varepsilon)\log(1/\delta) \sum_{q\in\widehat{\mathcal{S}}_\pi(B_\delta)} h_q\\ &\ge (1-\varepsilon)T_k^*(\mu)\log(1/\delta).\end{aligned}$$

After division by $\log(1/\delta)$, this says $(1-\varepsilon)T_k^*(\mu) \le b$, hence $k \le \bar{k}$ by definition. The argument also covers $k = 0$ since $T_0^* = 0$. If $\bar{k} = |\mathcal{G}|$, any output larger than $\bar{k}$ already contains a bad arm. Therefore

$$\begin{aligned}&\sup_{\pi\in\Pi_\delta(B_\delta)} \mathbb{P}_{\mu,\pi}\left(|\widehat{\mathcal{S}}_\pi(B_\delta)| > \overline{K}_\varepsilon(b)\right)\\ &\le \delta + 2|\mathcal{G}|M\delta^{\varepsilon^2/64} \longrightarrow 0.\end{aligned}$$

For $|\mathcal{G}| = 0$, the same conclusion follows directly from reliability. □